Mi:dm 2.0
Korea-centric Bilingual Language Models

Tech. Innovation Group, KT
midm-llm@kt.com
https://huggingface.co/K-intelligence

Abstract

We introduce Mi:dm 2.0, a bilingual large language model (LLM) specifically engineered to advance **KOREA-CENTRIC AI**. This model goes beyond Korean text processing by integrating the values, reasoning patterns, and commonsense knowledge inherent to Korean society, enabling nuanced understanding of cultural contexts, emotional subtleties, and real-world scenarios to generate reliable, culturally appropriate responses. To address limitations of existing LLMs—often caused by insufficient or low-quality Korean data and lack of cultural alignment—Mi:dm 2.0 emphasizes robust data quality through a comprehensive pipeline that includes proprietary data cleansing, high-quality synthetic data generation, strategic data mixing with curriculum learning, and a custom Korean-optimized tokenizer to improve efficiency and coverage. To realize this vision, we offer two complementary configurations: ***Mi:dm 2.0 Base*** (11.5B parameters), built with a Depth-up Scaling strategy for general-purpose use, and ***Mi:dm 2.0 Mini*** (2.3B parameters), optimized for resource-constrained environments and specialized tasks. Mi:dm 2.0 achieves state-of-the-art performance in Korean-specific benchmarks, with top-tier zero-shot results on KMMLU and strong results in internal evaluations across language, humanities, and social science tasks.

The Mi:dm 2.0 lineup is released under the MIT license supporting extensive research and commercial use. By offering these accessible and high-performance Korea-centric LLMs, KT aims to accelerate AI adoption across Korean industries, public services, and education, while strengthening the Korean AI developer community and laying the groundwork for the broader vision of K-intelligence. Our models are available via `https://huggingface.co/K-intelligence`. For technical inquiries, please contact `midm-llm@kt.com`.

1 Introduction

1.1 The Korea-centric AI, Mi:dm 2.0

KT (Korea Telecom) has developed Mi:dm 2.0 as an instruction-tuned language model that embodies what we call **KOREA-CENTRIC AI**. Korea-centric AI refers to a model that thoroughly internalizes the unique values, cognitive frameworks, and commonsense reasoning intrinsic to Korean society. It is not simply about processing and responding in Korean; it is about the profound understanding that reflects and respects the socio-cultural fabric of Korean norms and values.

The development of Mi:dm 2.0 begins with a critical observation: despite the release of numerous large language models (LLMs) supporting the Korean language, few are truly grounded in the realities of Korean society. We identify a pervasive problem where existing LLMs, often trained on insufficient or low-quality Korean datasets, exhibit limited linguistic performance and a noticeable gap in their alignment with Korean cultural sensibilities [1, 2]. This problem often leads to unnatural

or emotionally incongruent responses, or even responses in languages other than Korean, from the perspective of Korean users. To directly address these deficiencies, we conceive Mi:dm 2.0, aiming to establish a new standard for a genuine Korea-centric AI that deeply internalizes the unique values, cognitive frameworks, and common sense reasoning inherent to Korean society, moving beyond mere linguistic proficiency to embody cultural nuance.

Our journey begins with a robust data curation pipeline, which meticulously defines criteria for sourcing high-quality, culturally representative Korean text, complemented by innovative synthetic data generation techniques. Following this, we detail our sophisticated pre-training methodology, which ensures effective learning even with a reduced corpus size through careful data selection and strategic training. We then delve into our model optimization techniques, designed to deliver computational efficiency that surpasses comparable domestic and international models. Furthermore, we outline the post-training techniques employed to significantly enhance the model's ability to perform Korean socio-cultural reasoning and generate contextually appropriate responses. Finally, we provide comprehensive quantitative and qualitative evaluations of Mi:dm 2.0's performance, including benchmark comparisons that unequivocally demonstrate its advanced understanding of Korean language, culture, and society.

1.2 Mi:dm 2.0 Line-up: Base & Mini

We build Mi:dm 2.0 of two distinct parameter scales—11.5B and 2.3B—to meet diverse deployment needs. The Mi:dm 2.0 lineup is the result of extensive and systematic experimentation across a wide range of model architectures, parameter scales, and compression techniques. Through rigorous testing and iterative refinement, we identify configurations that deliver strong quantitative benchmark performance, robust language model evaluations, and reliable results from comprehensive human assessments using real-world scenario prompts.

Mi:dm 2.0 Base (11.5B) serves as a general-purpose foundation model, meticulously engineered to strike an optimal balance between scale and performance. Its development began with training an 8B-parameter model from scratch using KT's proprietary pre-training corpus. To further enhance its capabilities and reach the 11.5B scale, we then applied a Depth-up Scaling (DuS) strategy [3]. This innovative approach efficiently expands the model's depth without requiring complex architectural changes, enabling us to effectively leverage representations learned by the initial 8B model.

In contrast, Mi:dm 2.0 Mini (2.3B) provides a lighter and more compact alternative. It is specifically optimized for deployment on resource-constrained devices, prioritizing computational efficiency. Mi:dm 2.0 Mini emphasizes task specialization, with a particular focus on intent understanding and machine translation, making it highly efficient for specific applications where resources are limited. Table 1 provides detailed configurations for both models.

During training, we employ various optimization techniques, including parallelization and quantization, to maximize GPU resource efficiency. We also extend and customize the underlying training framework to meet Mi:dm's specific requirements. Mi:dm 2.0 delivers competitive or superior quality compared to open-source models of similar or larger scale while maintaining significantly lower computational overhead. We provide more detailed descriptions of model design and training methodology in Section 3.

Both Mi:dm 2.0 Base and Mi:dm 2.0 Mini are released under the MIT license, allowing extensive use for both research and commercial purposes. The models are available for download on Hugging Face and can be applied to a wide range of applications without restriction.

Looking ahead, we plan to expand the Mi:dm 2.0 lineup to support an even broader range of use cases, further solidifying its role as a foundation of K-*intelligence.* We are also committed to releasing our training code, serving environments, and other research artifacts to facilitate broader adoption and continuous improvement within the ecosystem. Through these efforts, we aim to accelerate AI transformation (AX) across Korean industry, public services, and education, and to make meaningful contributions to the Korean AI developer community with both the Mi:dm 2.0 Base and Mi:dm 2.0 Mini.

Models	Model Type	Model Size	Context Length	License
Midm 2.0 Base	Dense	11.5B	32K	MIT
Midm 2.0 Mini	Dense	2.3B	32K	MIT

Table 1: Mi:dm 2.0 line-up architecture and license

2 Data

This section details the methodologies for data construction and management, including domain classification, document filtering pipeline, and synthetic data generation strategies adopted in the development of Mi:dm 2.0. Recent LLMs are known to critically depend on large amounts of high-quality textual data to achieve robust performance [4, 5]. However, the Korean language poses a unique challenge due to the limited availability of high-quality and publicly accessible training data compared to English. Furthermore, the heterogeneous quality of Korean corpora available hinders stable and reliable performance. [6, 7]

To overcome the structural limitations of existing Korean training datasets, Mi:dm 2.0 is strategically designed and robustly trained from the earliest stages of pre-training corpus construction. We explicitly prioritize data quality over quantity to ensure the stable assimilation of both general knowledge and the nuanced Korean cultural and societal context. Despite the reduction in overall token volume, we focus on filtering for accurate, complete, and trustworthy documents.

For Mi:dm 2.0, we define high-quality data as documents that are contextually coherent, highly readable, non-toxic, and well-formed. To achieve this standard, our proprietary data cleansing pipeline strictly excludes documents that do not meet these criteria. Unlike English datasets, where quality selection is simpler due to the abundance of commercially available open-source corpora, applying the same stringent quality criteria to Korean text significantly reduces the total usable tokens. Nevertheless, we firmly maintain our focus on this data quality approach to enhance the model's overall representational power and generalization performance.

To address this, we deliberately generate high-fidelity synthetic data to supplement the limited volume of high-quality Korean corpora. This synthetic data, primarily grounded in organic (human-generated) content, is augmented by language models. Given the overarching challenge of limited availability and high acquisition costs for Korean datasets, many of which are web-based, synthetic data augmentation has proven to be particularly valuable. These include translating English corpora into Korean and generating textbook-style documents from topics and keywords extracted from existing bilingual corpora.

Furthermore, Mi:dm 2.0's data engineering pipeline incorporates data mixing and curriculum learning strategies. Concurrent with corpus filtering and augmentation, we establish a hierarchical domain taxonomy based on application-specific needs. This taxonomy then guides the alignment of training data distribution with the intended use cases of the model. To systematically monitor and manage dataset balance, we train a domain classifier across the entire corpus to quantify its distribution. This enables the model to identify underrepresented domains quantitatively and granularly. For domains lacking sufficient token density, we generate additional synthetic data via a feedback loop, enhancing both coverage and diversity of the training corpus.

To optimize data efficiency, our model utilizes a custom tokenizer specifically designed to capture the unique linguistic characteristics of Korean. Based on a precisely curated pre-training corpus, we design the Mi:dm 2.0 tokenizer to handle the morphological structure of the Korean language more effectively than traditional GPT-series tokenizers. This approach achieves higher token compression and significantly enhances computational efficiency during both training and inference.

The entire data pipeline for our model adheres to rigorous standards for data provenance, licensing, and compliance. All training data are sourced from open-source datasets or formal licensing agreements with third parties, ensuring legitimacy. We strictly exclude any data posing legal or ethical concerns, including unauthorized crawls or user-sensitive content. Notably, we exclude Personally Identifiable Information (PII) and proprietary customer data, thereby ensuring both data security and ethical responsibility throughout the model development process.

Category	Details
Language	English Korean Code Math Multi Language

Source	Organic
	  Web
	  Government
	  Book
	  News
	  Paper
	  Encyclopedia
	  Others
	Synthetic

Domain	Humanity STEM Applied Science Health & Food Life & Culture ETC
Expression Mode	Written Spoken
Stylistic Tone	Formal Informal

Table 2: Data Categorization: Categories and Details

2.1 Multidimensional and Hierarchical Data Classification Framework

Precise analysis and structured management of training data are essential for developing high-performance language models. However, the scarcity of detailed categorization for Korean language datasets in existing research significantly limits efforts to expand data coverage and interpret model performance [8]. To overcome these challenges, we define a novel data classification framework designed to support balanced data distribution and efficient training. It organizes data across multiple dimensions -including language, domain, source, and linguistic style -and is consistently applied throughout the entire data pipeline, from collection to training, as shown in Table 2. Hence, following the refinement and high-quality data selection, we classify data from multiple perspectives, as detailed below.

In detail, from the language perspective, our dataset is classified not only as multilingual text, such as Korean and English, but also as non-linguistic content, including mathematical expressions and source code. From a domain perspective, our dataset is categorized using an internally developed taxonomy designed to reflect both the thematic content and the intended application of each document. This taxonomy comprises six primary domains: Humanity, STEM, Applied Science, Health & Food, Life & Culture, and ETC, along with 20 mid-level subdomains that provide further granularity. Finally, for the data source, documents are broadly categorized as either organic or synthetic. Organic data consists of naturally occurring text derived from real-world human activity. This includes sources such as web pages, news articles, books, encyclopedias, government documents, academic papers, and other written materials. Conversely, synthetic data refers to text generated with augmentation techniques. This includes documents created through machine translation, document rewriting, and advanced methods such as Chain-of-Thought (CoT) generation [9]. Beyond these dimensions, we also classify texts according to their linguistic style and tone. Specifically, we categorize documents based on whether they predominantly feature written or spoken language characteristics and whether their tone is formal or informal.

Throughout the entire Mi:dm 2.0 training process, statistical information for each classification attribute is continuously managed across all training data subsets, from data collection to training, guided by this classification framework. Each training sample is tagged with up to five classification attributes in these dimensions. For instance, we would categorize a Korean-language web review of a children's book titled after the historical figure "King Sejong (세종대왕, Sejong Daewang)" as

"Korean" in the language dimension, as "History" subdomain within the "Humanity" domain, and as "Organic" in source. In another case, a Korean document generated by extracting the keyword "chlorophyll" from an English web article, rewriting the content in a textbook-like format, and translating it into Korean would be classified as "Korean" in language, "Biology" subdomain within the "STEM" domain, and as "Synthetic" in source. By maintaining comprehensive statistics across all classification axes, we can monitor data distribution, identify underrepresented categories, and strategically augment the dataset to ensure balance and diversity.

2.2 Composition of Data Sources

Among the multiple classification axes described in Section 2.1, data source serves as a critical criterion from the earliest stages of data collection. The goal is to ensure diversity within the pre-training corpus, enabling the language model to acquire expressive capabilities in various document styles and topics. This ultimately enhances its ability to generalize linguistically across different real-world contexts.

Accordingly, the pre-training corpus primarily consists of organic data —naturally occurring, human-authored text—to accurately reflect authentic language usage environments, as shown in Fig. 1. Approximately 85.7% of the total dataset is sourced from organic domains, such as web documents. This composition is carefully selected to enable the model to acquire high-fidelity knowledge of Korean-specific syntactic structures, discourse patterns, and informal expressions. Additionally, approximately 10% of the corpus comprises of open-source, organic data acquired from high-quality public datasets, such as AIHub[*] and the National Institute of the Korean Language (NIKL)[*]. These resources are comprised of administrative documents, transcribed spoken dialogues, and other publicly available linguistic assets, thereby enhancing the model's reliability, particularly in terms of standard language usage and compatibility with public benchmarks. A smaller subset—approximately 0.71%—is comprised of additional organic sources, including academic papers, books, government documents, and dictionaries. Although this category constitutes a relatively minor portion of the corpus, it significantly contributes to the performance of the model by expanding its capacity to understand high-density informational text, formal language, and conceptually coherent content. This data composition strategy extends beyond exclusively ensuring topical variety. It aims to provide the model with a broad range of linguistic contexts found in real-world Korean use, helping it become more natural, robust, and contextually aware.

To enable our model to effectively embody 'Korea-centric AI,' the data acquisition and corpus construction strategy prioritized critically maximizing the linguistic and topical diversity of its Korean-language organic dataset. We employ a systematic corpus construction methodology that integrates both public collections and licensed acquisitions. This approach yields a comprehensive array of resources, including Korean literary works, modern historical records (e.g., news articles), public documents (e.g., legal texts and dictionaries), and structured databases specifically curated for Korean cultural heritage.

The major sources of organic data derive from Common Crawl (CC)[*], Hugging Face[*], AIHub, and the NIKL. News articles, books, and dictionaries are obtained via formal licensing agreements and subsequently filtered according to internal quality standards. For the English web corpus, we leverage open-source corpora with pre-annotated document-level quality indicators. In contrast, Korean web data is sourced from CC-based corpora and processed through an internally developed filtering pipeline to ensure suitability. Furthermore, data acquired from AIHub and NIKL undergo explicit permissioning and rigorous curation to ascertain their applicability for commercial deployment.

Synthetic data accounts for approximately 14% of the entire training dataset. Its primary purpose is to compensate for domain-specific data scarcity in organic Korean datasets, specifically to augment coverage in underrepresented areas. Furthermore, synthetic data generation was employed to enhance knowledge about Korea that was insufficient in the original organic data source. For English, given the abundance of high-quality open datasets, we selectively integrate publicly available synthetic corpora. Conversely, we produce Korean synthetic data via custom-built generation pipelines. The

[*] https://www.aihub.or.kr
[*] https://www.korean.go.kr
[*] https://commoncrawl.org
[*] https://huggingface.co

construction of synthetic data in our model leverages both publicly available research methodologies and proprietary augmentation techniques developed in-house [10, 11]. These synthetic data generation strategies are further elaborated in Section 2.4.

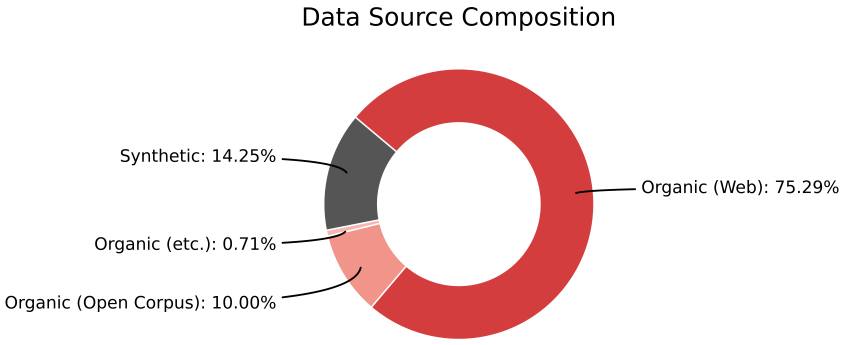

Figure 1: Distribution of Dataset by Source Type (Organic vs. Synthetic)

2.3 High-Quality Data Filtering and Refinement Pipeline

Mi:dm 2.0 implements an intentionally developed quality control pipeline for data selection and pre-processing, ensuring that the resulting corpus is optimally suited for next-token prediction training. The filtering criteria are defined to construct token sequences that are both coherent and learnable, minimizing interference during model training.

From a pre-training perspective, high-quality data is defined as text that meets the following conditions:

1) Data should maintain consistent textual coherence, devoid of disruptive special characters or grammatical malformations.
2) Data that satisfies 1) and should be comprised of highly readable and well-formed complete sentences, adhering to standards of linguistic completeness.
3) Data should be devoid of harmful content and free from any personally identifiable information that could compromise privacy.

These quality standards are applied throughout the data preparation pipeline, particularly during the pre-training phase, where the generalization capability of the model is shaped. By minimizing the incidence of noisy or irrelevant tokens, the model can learn more stable token distributions and more effectively internalize linguistic patterns and knowledge.

To achieve this, first our model employs multi-stage data filtering and refinement pipeline that is both language- and source-specific, with a particular emphasis on Korean-language data to mitigate the relative scarcity of high-quality Korean tokens. As illustrated in Fig. 2, the primary component of

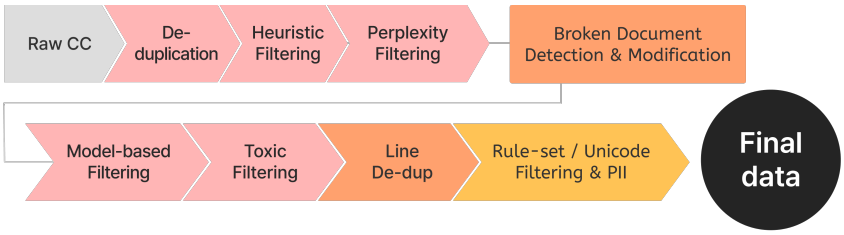

Figure 2: Multi-Stage Filtering and Refinement Pipeline for Korean Web Documents

this effort is an internally developed, 8-stage filtering pipeline for large-scale Korean web data. Web

corpora—especially those derived from CC —contain documents with format corruption, low-quality text, harmful or biased content, or PII, rendering raw web data unsuitable for direct use. Our model addresses this by applying a rigorous, sequential filtering strategy specifically tailored to Korean web content. Each stage of the pipeline incrementally refines the dataset, isolating documents of genuine learning value:

1) **Document De-duplication**: Redundant documents are removed based on cosine similarity over TF-IDF vectors.
2) **Heuristic Filtering**: Documents containing hashtags, excessive ellipses, or abnormal punctuation are filtered using handcrafted rules inspired by prior work [3].
3) **Perplexity Filtering**: Documents exhibiting abnormal n-gram perplexity are removed as likely being low-quality or incoherent.
4) **Broken Document Detection and Correction**: Unicode corruption and broken character sequences are detected and corrected.
5) **Model-Based Quality Filtering**: An ensemble of binary classifiers is trained using annotated examples of high- and low-quality documents. An ensemble of binary classifiers is used, comprising one classifier trained on general quality criteria and another trained on educational quality criteria inspired by prior studies [12, 13].
6) **Toxic Content Filtering**: A binary classifier trained on KT's proprietary Korean toxicity and bias taxonomy is used to eliminate harmful or offensive content.
7) **Line-Level De-duplication**: Within a document, repetitive lines or paragraphs are removed to reduce redundancy.
8) **Final Rule-Based Refinement and PII Anonymization**: Final cleanup includes Korean-specific formatting fixes, normalization of invisible Unicode tokens, and removal or anonymization of any detected personal information.

For the subsequent step, we design a source-specific refinement pipeline for non-common Crawl Korean datasets, such as those securely acquired from books, encyclopedias, academic papers, expert knowledge databases, and licensed news articles. This pipeline consists of source-specific refinement modules for each domain and explicitly reflects the unique characteristics of Korean language data.

For instance, the news article refinement module incorporates rules for removing strings that are irrelevant to the core content of individual articles or disrupt the overall context. The bylines (reporter names and email addresses) at the end of Korean articles or image captions remaining within the main body after image removal are removed. Additionally, rules applicable to Korean data, such as eliminating string patterns like '[속보]' or '상보', which are uniquely found in domestic Korean online news headlines, are also incorporated.

As another example, the Korean court judgment refinement module applies rules to redact PII and extract only the critical content from the unique formatting of domestic court judgments while preserving the structural and semantic content necessary for learning formal legal language.

Documents refined through these source-specific rules are considered comparable in quality to those that have passed harmful content filtering in the Korean web data pipeline. Subsequently, they undergo final steps such as deduplication and PII anonymization before being used in the training dataset.

Finally, for English, code, and mathematical content, high-quality, commercially suitable public datasets are selected and rigorously assessed during acquisition through manual inspection and sample-based qualitative analysis. Deduplication and final normalization are performed to ensure alignment with the quality standards of the Korean dataset.

2.4 Synthetic Data Generation

It is essential to acquire high-quality data, encompassing diverse knowledge and linguistic expressions across a wide range of domains, for our model to enable complex reasoning and conceptual understanding. In particular, being Korea-centric AI requires gathering a collection of Korean-language datasets that reflect sufficient diversity and representativeness. However, in practice, the availability

of open-access Korean-language corpora is significantly lower than that of English, and a large portion of the available data is concentrated in web sources, which often contain low-quality content. Furthermore, the majority of Korean-language data is disproportionately biased towards humanities and social sciences, creating additional challenges for domain diversity.

These structural limitations are empirically observed in the data distribution statistics collected after sourcing from various domains. Fig. 3 illustrates the token distribution of Mi:dm 2.0's pretraining corpus across data sources and domains. The light blue bars, which represent tokens from the humanities and social sciences, account for a disproportionately large portion of the dataset. In contrast, domains such as applied sciences (APSC), arts (ARTS), and culture (CULT) are severely underrepresented. Notably, applied sciences constitute only 0.1% of the total token count, highlighting a clear example of domain imbalance within the current data ecosystem. Such an imbalance has the potential to affect the model's expressiveness and reasoning capabilities in domain-specific applications.

To mitigate such biases, Mi:dm 2.0 strategically incorporates high-quality synthetic data during the pre-training phase. The generation pipeline extends simple translation-based augmentation, specifically designed to simulate reasoning structures and compositional understanding by producing textbook-style explanatory passages, logically structured reasoning chains, and diverse document types tailored to specific learning objectives. All synthetic data undergoes the same rigorous post-processing as the organic corpus, including Unicode normalization, PII filtering, and deduplication, to ensure consistency in quality before being included in the final training dataset.

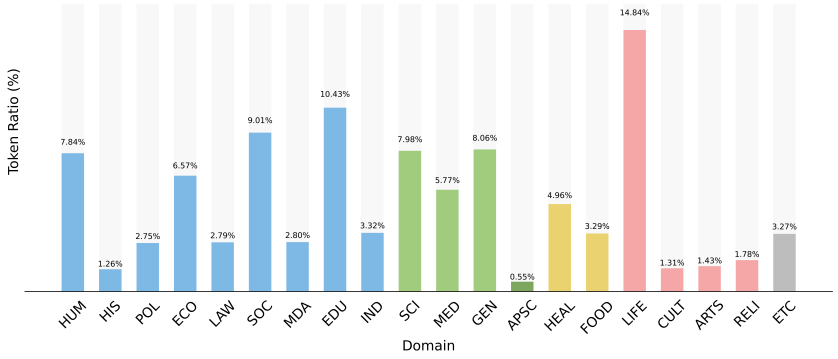

Figure 3: Domain-wise token distribution of Mi:dm 2.0 pretraining corpus

2.4.1 Corpus Construction for Domain Balancing in Korean-Language Data

An analysis of Mi:dm 2.0's pre-training data distribution reveals that domains, particularly those underrepresented in the Korean corpus collected from public datasets, are computationally intensive fields such as STEM (science, technology, engineering, and mathematics) and economics. Such imbalance in the dataset is reflected in early benchmark evaluations during model development, where our model consistently underperformed in disciplines that demand high-level reasoning and domain-specific knowledge, such as physics, chemistry, biology, mathematics, computer science, and economics.

To address this structural bias, we systemically collect high-reliability open-source materials to serve as seed data for domain-targeted synthetic augmentation. These seed corpora are carefully selected to introduce and reinforce conceptual understanding and problem-solving capabilities in previously underrepresented domains. Leveraging insights from prior work [11, 10], we tailor prompt structures for each domain. Specifically, core concepts derived from seed documents are converted into high-quality Korean language instructional texts, utilizing various formats, including textbook-style explanations and scenario-driven narratives. The synthetic documents are generated at varying levels of difficulty and designed for diverse reader profiles, thereby enriching the corpus in domains lacking sufficient coverage.

2.4.2 Reconstruction and Augmentation of Non-Selected Korean Web Documents

The portion of Korean language in the CC corpus integrated into our model's pre-training dataset stems from a widely used open-source resource for large-scale language model training. Despite applying a rigorous filtering pipeline to extract only a small subset of high-quality documents, the CC corpus still holds the largest portion of the Korean-language dataset used in our model.

However, this dataset is inherently noisy, with a significant portion of its documents being of low quality. In fact, over 80% of the initially collected raw CC data is subsequently excluded during the data filtering process. This highlights a structural limitation of CC: while it provides broad coverage, the usable token yield relative to curation effort remains low.

To overcome this inefficiency and increase the number of usable Korean-language tokens for pre-training, we develop a rewriting-based synthetic reconstruction strategy for a portion of the filtered-out CC documents. Manual inspection of rejected samples reveals that some documents—although initially discarded due to poor formatting or content—can be transformed into high-quality training material if their core topics and sentence structures are properly reorganized. Given the lack of consistent structural patterns, these documents cannot be effectively recovered using rule-based methods. Consequently, we developed a generative rewriting pipeline specifically for Korean CC documents.

This rewriting process is composed of two prompt stages. In the first stage, the topic analysis module extracts metadata, including the central topic of the document and relevant paragraph indices. This enables the filtering of irrelevant fragments within documents—such as image captions, template-based advertisements, or copyright notices—often found as short, extraneous sentences. This stage also identifies and separates cases where a single document actually contains multiple unrelated articles concatenated together. In the second stage, based on the topic structure extracted earlier, the pipeline generates excerpted and rewritten documents, focusing only on topic-relevant content. The rewriting model synthesizes new documents that preserve the central meaning of the original while eliminating noise and improving coherence.

Finally, all reconstructed documents are passed through the same web corpus filtering pipeline used for the original CC documents. This ensures that any rewritten documents containing harmful content, biased language, or incoherent structure are excluded. Only documents that met the same high-quality criteria as the original filtered set are included in the final pre-training corpus.

2.4.3 Structural Augmentation of English Web Documents

Our strategy synthetically enhances the structural diversity and complexity of Korean language web data, which is primarily derived from CC. Despite CC's known limitations—namely, its high proportion of low-quality content—it remains a valuable corpus as it closely reflects modern language use in real-world human contexts.

In English-speaking communities, several high-quality web corpora curated from CC have been released, accompanied by research on large-scale web data filtering methods [13, 14]. These efforts have provided models with rich, well-structured input across various formats.

In contrast, the Korean CC corpus tends to be limited in both structural diversity and topical breadth, being heavily skewed toward specific formats such as news articles, blogs, and online community posts. As a result, it lacks the structural richness and stylistic complexity observed in English corpora. Notably, Korean CC contains relatively few examples of long-form structured documents or intent-driven formats (e.g., summarization, QA, translation), which are indispensable for training models on higher-order reasoning and downstream tasks such as question answering.

To mitigate these limitations, our model integrates a cross-lingual synthetic augmentation strategy, where unused English web samples are rewritten into Korean texts during pre-training. These rewritten documents are not direct translations, but content-preserving rewrites into natural Korean formats that differed structurally from the original web style. This approach helps avoid typical translation errors—such as the literal rendering of idiomatic expressions or the misinterpretation of domain-specific terminology—often seen in naïve machine translation.

For example, the content of an English web document can be transformed into the style of a Korean university entrance exam question in the "Speaking and Writing" section. In this process, the original content is transformed into a coherent Korean passage, accompanied by reading comprehension

questions and answer sets. This enables the model to learn both richly composed texts and QA-style supervision. In practice, the majority of synthetic QA documents contributed solely the passage portion to the final pre-training corpus, while a select subset retained the QA pairs. The QA dataset included in the corpus is further filtered through CoT-based verification, and only samples with verified correct answers are retained.

Meanwhile, we opt not to integrate any additional rewriting on previously synthesized data to circumvent the risks of semantic drift, factual inconsistency, or unintended duplication. This decision preserves the integrity and quality of the final corpus.

2.4.4 Structured Long Chain-of-Thought Data for Math and Code Reasoning

We construct the LongCoT dataset [9] to provide synthetic problem-solving sequences that explicitly model multi-step reasoning. Each math or code example includes a clear, logically structured solution path designed to help the model learn the reasoning patterns required for complex tasks.

This resource not only supplements the limited availability of Korean-language data but also improves the model's ability to learn reasoning in Korean. All solutions and explanations are written primarily in Korean to help the model develop native-level logical reasoning for structured problem domains.

The final data is formatted into pre-training-ready text segments and integrated directly into the Mi:dm 2.0 training set. By exposing the model to high-quality reasoning demonstrations early on, this approach supports stronger performance in math, programming, and structured question-answering tasks.

3 Pre-training

This section outlines the model lineup of Mi:dm 2.0 and the corresponding pre-training strategies applied to each model. Mi:dm 2.0 achieves efficient pre-training and strong performance in Korean understanding and generation tasks by leveraging a small, highly curated Korean-language dataset with limited computational resources. This approach demonstrates the feasibility of developing competitive language models even in resource-constrained environments.

To introduce the pre-training methodology for each model in detail, Section 3.1 provides an overview of the model lineup expansion process. This is followed by subsections describing the pretraining techniques for each model variant: Section 3.1.1 covers Mi:dm 2.0 Base and Section 3.1.2 discusses Mi:dm 2.0 Mini. Section 3.1.3 then presents the sequence length extension techniques applied during training. Finally, details the computational cost optimization strategies that enable efficient large-scale training.

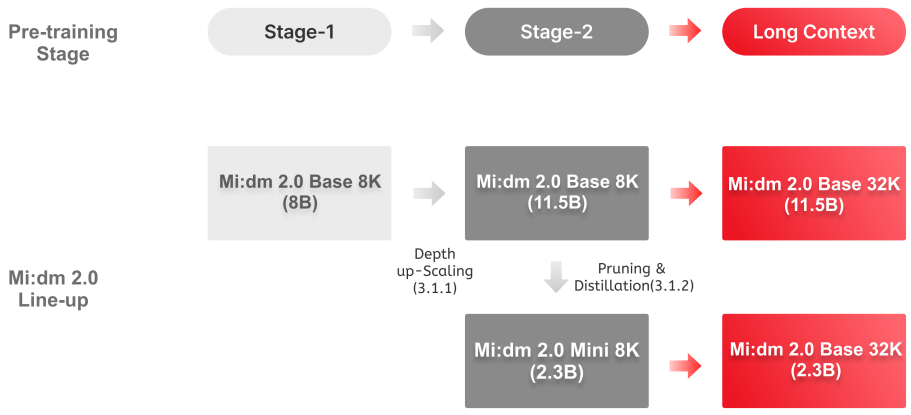

Figure 4: Mi:dm 2.0 Model Lineup and Pre-training Pipeline

3.1 Model Architecture

All Mi:dm 2.0 lineups are based on a Transformer decoder-only architecture [15]. The number of layers in this architecture is adjusted to optimize training efficiency and performance across various model sizes and applications.

As illustrated in Fig. 4, the expansion of the model architecture employs a three-stage training procedure. In the initial pre-training phase (Stage-1), an 8B parameters Mi:dm 2.0 Base is trained from scratch. This foundation model, entirely developed in-house from architectural design to training, does not leverage existing open-source weights. The Stage-1 model comprises of 32 layers of dense Transformer decoders and processes input sequences up to approximately 8,000 tokens. The objective of this stage is to establish a foundation model that can acquire general language abilities and extensive domain knowledge. To achieve this, the training dataset for Stage-1 is curated to cover a wide range of topics and domains.

In the second stage (Stage-2), the focus shifts to scaling up the model using the checkpoint from Stage-1. The outcome of this stage is the finalized Mi:dm 2.0 Base. Depth up-Scaling (DuS) technique [3] is applied to increase the number of transformer decoder layers from 32 to 48, resulting in a parameter count of 11.5B. Although the volume of the training data in Stage-2 is smaller than that of Stage-1, it is composed of ultra-high-quality data to enhance the model's ability to generate refined, task-specific responses.

Based on the Mi:dm 2.0 Base, after completing its second training phase, Mi:dm 2.0 Mini is trained with a reduced size of 2.3B parameters through quantization. To preserve the core knowledge of the Mi:dm 2.0 Base while reducing the size, both width-based pruning techniques [16, 17] and multi-stage knowledge distillation [18, 19] are employed. The resulting Mi:dm 2.0 Mini is optimized for on-device deployment, enabling efficient inference in resource-constrained environments.

In the final stage, long-context training is applied to both models. This involves extending the base frequency in Rotary Position Embedding (RoPE) [20] to process longer input sequences. Whereas the original Mi:dm 2.0 Base and Mi:dm 2.0 Mini could only handle sequences up to approximately 8,000 tokens, long-context training extends their maximum input token length to approximately 32,000 tokens, allowing for more effective processing of long documents. The details of each Mi:dm architecture are summarized in Table 3.

Specification	**Mi:dm 2.0 Mini**	**Mi:dm 2.0 Base**
Number of Parameters	2.3B	11.5B
Hidden size	1,792	4,096
Number of layers	48	48
Activation function	SiLU	SiLU
Feedforward Dimension	4,608	14,336
Attention type	GQA	GQA
Number of attention heads	32	32
Head size	128	128
Context length	32,768	32,768
Positional Embeddings	RoPE(θ=8,000,000)	RoPE(θ=8,000,000)
Vocab size	131,392	131,384
Tied word embedding	True	False

Table 3: Detailed architectural specifications of the Mi:dm 2.0 models. Both variants share key design choices such as the number of layers, attention mechanism (GQA: Grouped Query Attention [21]), and activation function (SiLU [22]), while differing in hidden size, parameter count, and feedforward dimensions. The table also reports positional embedding configuration (RoPE) and vocabulary size.

3.1.1 Mi:dm 2.0 Base: Depth Up-Scaling

We design Mi:dm 2.0 Base for robust performance across diverse application environments, even with limited computational resources. This is achieved through DuS, a training methodology that systematically duplicates specific layers from the base model and stacks them on top of existing layers to increase the model's depth.

The efficacy of DuS depends on strategically choosing layers with strong representational capacity. Effectively choosing layers with strong representational capacity allows DuS to maximize its advantages, enhancing the expressiveness and performance of the model while efficiently reusing resources from the initial training phase.

For Mi:dm 2.0 Base, we adopt a quantitative methodology, as proposed in [23], to determine which layers to replicate. This approach measures the change in embedding representations before and after each layer by calculating the cosine similarity between them. Layers with higher cosine similarity scores are selected for duplication under the assumption that they stably preserve the input information with minimal degradation during the training.

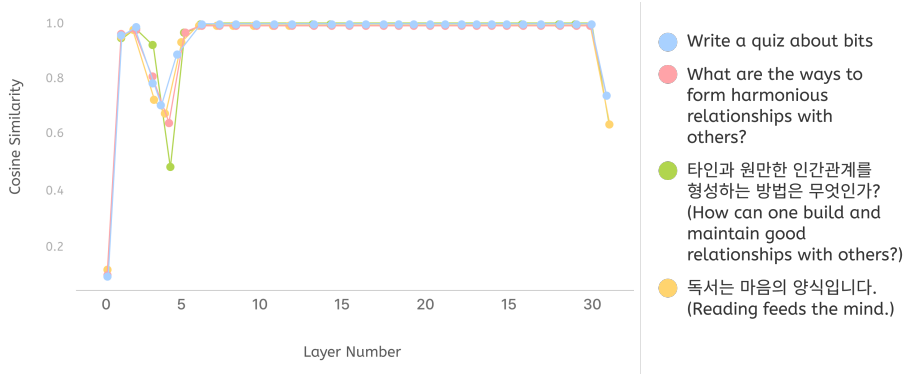

Figure 5: Analysis of Layer-wise Embedding Changes Based on Cosine Similarity

Applying this method, we analyze the embedding variation across all layers of the initial 8B-scale model used for Mi:dm 2.0 Base expansion. As shown in Fig. 5, we compute the cosine similarity between pre- and post-layer embeddings for each token in the input sentence and then average the scores per layer. This analysis reveals that the embedding changes stabilize after the 5th layer, with layers between the 7th and 29th showing near-perfect similarity (close to 1.0), indicating minimal information loss.

We confirm these results through Fig. 5 and additional internal experiments on other language models with similar architectures, which show consistent patterns as reported in [23]. Based on these findings, we select the continuous range from the 7th to the 29th layers for duplication, extending the total number of layers to 48 in Mi:dm 2.0 Base.

To verify the effectiveness of this architectural expansion, we evaluate the performance of the expanded model in its initial state. As shown in Table 4, the structure-extended but untrained Base-init model demonstrates comparable benchmark performance to the original Mi:dm 2.0 Base-8B, suggesting that our cosine similarity-based layer selection method successfully achieves effective structural scaling.

After this layer expansion step, Mi:dm 2.0 Base undergoes continual pre-training with thoroughly refined high-quality data to further improve its linguistic expressiveness and general domain understanding. This continual pre-training is structured in two stages: the first stage aims to ensure stable convergence immediately after expansion, while the second stage focuses on strengthening Korean and STEM domains using curated Korean data and custom-generated synthetic data. The training phase employs a Warmup-Stable-Decay scheduler [24] with a peak learning rate of 3e-4, decaying linearly to 0.0 during the final 10% of training. As shown in Table 4, the results demonstrate improved performance across various domains, including Korean and mathematics.

3.1.2 Mi:dm 2.0 Mini: Pruning and Distillation

Mi:dm 2.0 Mini is a lightweight variant derived from Mi:dm 2.0 Base, designed to run on low-resource environments such as on-device deployments or low-spec GPUs. To retain the knowledge acquired by Mi:dm 2.0 Base while reducing the model size, Mi:dm 2.0 Mini undergoes two stages of pruning and distillation, as illustrated in Fig. 6. In the first stage, we apply width pruning to Mi:dm 2.0 Base to produce an intermediate model of approximately 5B parameters, which we refer to as Mi:dm

Model	MMLU	HellaSwag	KMMLU	HAERAE	GSM8K
Mi:dm 2.0 Base-8B	51.94	74.98	29.36	56.18	14.48
Mi:dm 2.0 Base-init	52.05	74.43	29.60	56.82	12.89
Mi:dm 2.0 Base*	**62.61**	**79.36**	**47.67**	**78.19**	**49.20**

Table 4: 5-shot performance results on five evaluation benchmarks. **Mi:dm 2.0 Base-8B** refers to the base model before DuS is applied to Mi:dm 2.0 Base-init. **Mi:dm 2.0 Base-init** is the model after DuS without additional training. **Mi:dm 2.0 Base*** represents an intermediate checkpoint of Mi:dm 2.0 Base-init, trained on 100B tokens.

2.0 Base-half. This intermediate model then undergoes knowledge distillation, with Mi:dm 2.0 Base acting as the teacher model, guiding the student model to mimicking its output. In the second stage, we apply further width pruning to this intermediate model. We also adopt a weight-sharing structure for the word embedding to finalize Mi:dm 2.0 Mini's architecture. During this second distillation stage, we use both the intermediate model (Mi:dm 2.0 Base-half) and the Mi:dm 2.0 Base as teacher models, sequentially training the student model to mimic their outputs.

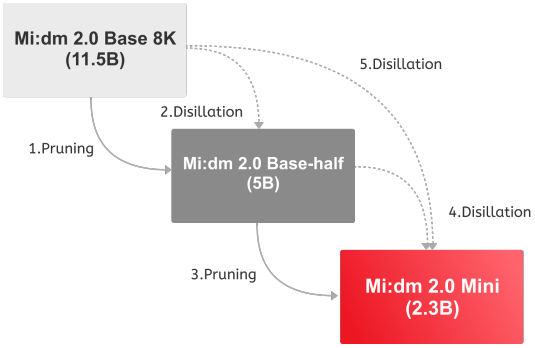

Figure 6: Pruning and Distillation Workflow of Mi:dm 2.0 Mini

This multi-stage knowledge distillation approach addresses challenges that arise when there is a large capacity gap between Mi:dm 2.0 Base and Mi:dm 2.0 Mini [19, 17]. When the disparity between the teacher and the student model size is significant, single-stage distillation often fails to effectively transfer knowledge and can destabilize the training process. Additionally, repeatedly relying on a large teacher model increases the overall computational cost[18, 19].

To mitigate these challenges and strike a balance between performance and efficiency, we introduce an assistant teacher model—approximately half the size of Mi:dm 2.0 Base—as an intermediary in our pruning-distillation pipeline [18, 16]. This intermediate-sized model serves as a critical mediator, enabling more stable and effective knowledge transfer while optimizing both learning performance and computational efficiency.

Furthermore, our width pruning strategy preserves the depth introduced by [3] during Mi:dm 2.0 Base construction while effectively reducing the overall parameter count. In the subsequent knowledge distillation process, the student model's initialization and structural design are critical for effectively absorbing knowledge from the teacher. Directly pruning the teacher model for student initialization can cause over-reliance on the original architecture and limit generalization. Prior studies have shown that, for equivalent parameter budgets, deeper models typically outperform wider ones [18, 25]. Guided by this insight, we adopt a strategy to reduce the width rather than depth when defining the intermediate and final model architectures.

To validate the effectiveness of our pruning-distillation strategy for Mi:dm 2.0 Mini, we conduct proxy model experiments following methodologies in [26, 23, 19]. Specifically, we use a 1.8B-parameter model sharing the same architecture as Mi:dm 2.0 Base as a proxy. We compare a scratch-trained version (1.8B-scratch) with a pruned-and-distilled version (1.8B-distill) across major benchmarks, as shown in Table 5. Results demonstrate that the distilled model achieves superior performance on most benchmarks while reducing computational cost by approximately 4.5 times lower compared to the scratch-trained baseline.

In both pruning stages, width pruning primarily targets the model's hidden dimensions and MLP dimensions. For post-pruning calibration, we sample 1,024 examples from Stage-2 pre-training data of Mi:dm 2.0 Base. The same sampling approach is applied to construct the knowledge distillation dataset. Distillation training uses a peak learning rate of 1e-4 with warm-up and cosine decay scheduling, following details described in [19].

Type	**MMLU** (acc, 5-shots)	**AGIEval** (acc, 5-shots)	**Winogrande** (acc, 5-shots)	**NQ** (EM, 64-shots)
1.8B-scratch	26.92	18.59	64.25	9.89
1.8B-distill	42.49	19.55	65.27	9.36
Type	**TriviaQA** (EM, 64-shots)	**KMMLU** (EM, 5-shots)	**HAERAE** (acc-norm, 3-shots)	**GSM8K** (EM, 5-shots)
1.8B-scratch	33.26	29.79	23.01	12.13
1.8B-distill	34.59	32.35	52.80	32.22

Table 5: Validation results of distillation-based lightweighting using proxy models. The 1.8B-scratch model is trained from scratch, whereas the 1.8B-distill model is distilled from a larger model. *acc* and *EM* denote Accuracy and Exact Match, respectively.

3.1.3 Long-context Extension

To enable Mi:dm 2.0 to handle long input sequences, we introduce an additional long-context training phase at the final stage of pre-training. This phase extends the model's maximum input context length from 8,192 tokens to 32,768 tokens.

Following insights from [20] on the relationship between base frequency in positional encoding and context length, we adjust Mi:dm 2.0's base frequency from 10,000 to 8 million. Additionally, as highlighted in [27], training with even longer sequences improves performance when targeting context lengths of 32K tokens. Therefore, long-context training uses data with input lengths of up to approximately 65,000 tokens.

Before this final training phase, we conduct experiments to validate data mixing strategies for long-context learning and to mitigate catastrophic forgetting as input length increases. Based on these results, the training dataset is primarily composed of sequences up to 65,000 tokens, with a small proportion of shorter data packed to match this length. Long context training is performed for 2,000 steps using a fixed learning rate of 1e-5 without a separate warm-up phase.

3.2 Training Costs

Mi:dm 2.0's large-scale pre-training is conducted on a high-performance GPU cluster built with Microsoft Azure CycleCloud [28]. This infrastructure is optimized for large-scale computations required in language model training, providing efficient resource management and flexible scalability to support long-term pre-training.

The computing cost at each training stage for each model is summarized in Table 6. All metrics are reported based on a cluster of NVIDIA H100 GPUs. Floating-point operations (FLOPs) are calculated based on the number of model parameters, sequence length, training steps, and batch size.

Model Type (Size)	**Total Amount of Computation (FLOPs)**
Mi:dm 2.0 Mini (2.3B)	4.57×10^{21}
Mi:dm 2.0 Base (11.5B)	1.74×10^{23}

Table 6: Training costs for Mi:dm 2.0. For Mi:dm 2.0 Mini, FLOPs are calculated based only on the student model, excluding the cost of the teacher model.

4 Post-training

4.1 Overview

Pre-trained LLMs possess a wide range of linguistic understanding and generation capabilities based on vast text corpora. However, to achieve the level of reliability required for real-world applications—such as precise instruction-following, logical reasoning, utilization of up-to-date information, tool use, safety, and long-context handling—further fine-tuning is essential.

Accordingly, the post-training process of Mi:dm 2.0 is designed to systematically enhance six key capabilities critical for maximizing utility and trustworthiness in actual service environments:

1) **Instruction-Following (IF)**: IF refers to the ability to interpret an instruction accurately and generate responses that match the requested content type, format, length, and structure. Since user queries and demands vary widely in real-world scenarios, strict adherence to instructions is necessary to ensure appropriate information delivery and service quality.
2) **Reasoning**: Reasoning refers to the ability to solve complex problems through logical and mathematical thinking, including multi-step operations. These abilities are essential for practical applications and determine the model's utility in real-world tasks.
3) **Retrieval-Augmented Generation (RAG)**: RAG refers to the capability to retrieve external knowledge, documents, or databases in real time and generate responses based on accurate and up-to-date information. This minimizes hallucinations and supports trustworthy decision-making in professional settings.
4) **Agent Ability**: Agent ability refers to the capacity to call various tools or APIs via designated interfaces to perform real-world tasks. Beyond simple Q&A, this is a core competency for service-oriented AI that handles complex scenarios such as scheduling or code execution.
5) **Safety**: Safety refers to the ability to ensure social and ethical responsibility, including harmlessness, bias mitigation, and privacy protection. Strengthening this area is vital to safeguarding users and organizations from harmful content, misinformation, or data leakage during deployment.
6) **Long Context Handling**: Long context handling refers to the ability to retain and consistently reference important information across long documents, conversations, or code. This ensures coherent and accurate responses in complex tasks such as summarization of lengthy materials.

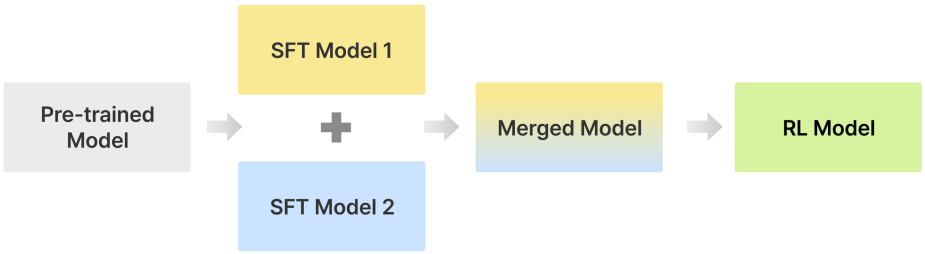

Figure 7: Post-training process of Mi:dm 2.0

To embed these capabilities, Mi:dm 2.0's post-training consists of structured alignment and specialization strategy to meet real-world demands. The overall post-training process is illustrated in **??** and is consisted of the following steps: supervised fine-tuning (SFT) on specific tasks, weight merging of multiple SFT models, and final preference optimization through reinforcement learning (RL).

- **SFT**: This step focuses on training the model to to balance generality and specificity, enabling broad responsiveness to user queries.

- **Weight Merging**: This step integrates strengths and features of multiple SFT models via weight merging [29], allowing diverse capabilities acquired through different data and training strategies to be unified in a single model.
- **RL**: This step enhances the model via online/offline reinforcement learning based on human or AI preferences, allowing the model to generate responses that better align with desirable qualities in both content and form.

This multi-stage strategy is designed to create synergies across four key aspects: (1) balancing specificity and generality, (2) efficiently integrating individual strengths, (3) improving training efficiency, and (4) aligning with core preferences.

Furthermore, unlike pre-training, post-training must incorporate data structures that meet real service environment requirements. While pre-training typically uses unstructured or single-turn corpora, real conversational services demand multi-turn interactions and role-based dialogue.

Therefore, Mi:dm 2.0 adopts the same chat template format used in LLaMA 4 [30] during post-training to reflect multi-turn interaction and role separation. Each utterance is clearly marked by roles such as `System`, `User`, `Assistant`, or `Tool`. Data is stored as structured multi-turn dialogues rather than simple corpora, and this structure is maintained throughout training and inference.

This structural approach encourages the model to learn realistic conversational scenarios, understand user intent, follow system instructions, and call appropriate tools for each contextual situations. It also improves the model's ability to maintain consistency, track information, and reference context according to each role.

4.2 Training Strategy

4.2.1 Supervised Fine-Tuning

SFT is the starting point of Mi:dm 2.0's post-training and plays a key role in equipping the model with real-world capabilities. From a mixture of datasets with diverse objectives and characteristics, we allocate fixed proportions to specific capabilities. Using this curated dataset, multiple SFT models are then trained using supervised learning. After training, their weights are merged so that the diverse capabilities acquired by each model are integrated into a single model. This allows the strengths of each model to be combined effectively.

To ensure that the model acquires a well-balanced set of abilities, each training batch is constructed by mixing samples across different capabilities. Definitions of these capabilities and data construction methods are described in 4.3. To encourage cross-lingual transfer from English (a high-resource language) to Korean (a low-resource language), a portion of translated data is also included in this phase. This batch design aims not only to specialize the model in specific areas but also to broaden its ability to handle a wide range of tasks and complex user demands in real-world service environments. In particular, the complementary nature of English and Korean is fully leveraged, focusing on using high-quality English data to improve Korean performance in a balanced manner. The proportions of Korean and English datasets used in Mi:dm 2.0 SFT training are shown in Fig. 8 and Fig. 9.

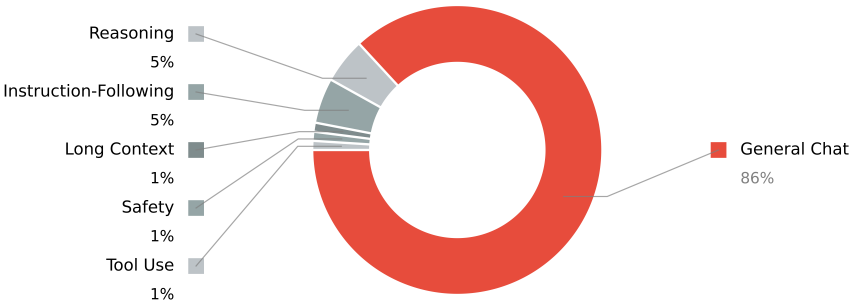

Figure 8: SFT dataset composition ratio (Korean)

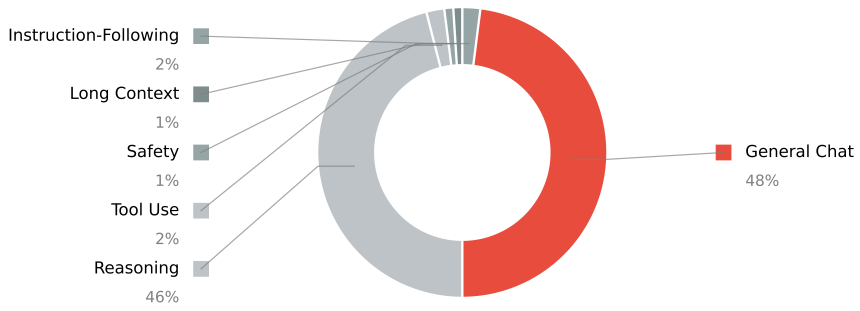

Figure 9: SFT dataset composition ratio (English)

During the SFT phase, the final model is obtained using the training configuration that yields both the most stable and highest-performing results across various experiments. Key considerations include GPU utilization, memory efficiency, and training speed in large-scale multi-node environments. Accordingly, hyperparameters were repeatedly tuned to derive optimal values. The training pipeline incorporates techniques such as data/tensor/pipeline parallelism, which are well-suited for large language model training. The settings for final model merging were also selected based on experimental results comparing multiple weight merge ratios.

4.2.2 Preference Optimization

Preference optimization focuses not just on generating correct answers, but also on producing responses that align with actual user experience and expectations. While SFT enables the model to learn patterns and explicit answers from data, real-world applications require the model to generate responses that reflect user preferences, expectations, and intent.

To this end, Mi:dm 2.0 incorporates preference-based training using datasets labeled with human or AI preferences. This approach allows the model to prioritize generating responses that align with key attributes such as safety, reliability, and usefulness. As a result, the model can produce more refined and appropriate responses compared to the baseline SFT model, significantly enhancing its usability, trustworthiness, and user satisfaction.

Fig. 10 shows the dataset composition used in Mi:dm 2.0's preference optimization stage. Like the SFT stage, this training phase also employs the state-of-the-art parallelization techniques also conducted in the same large-scale training environment.

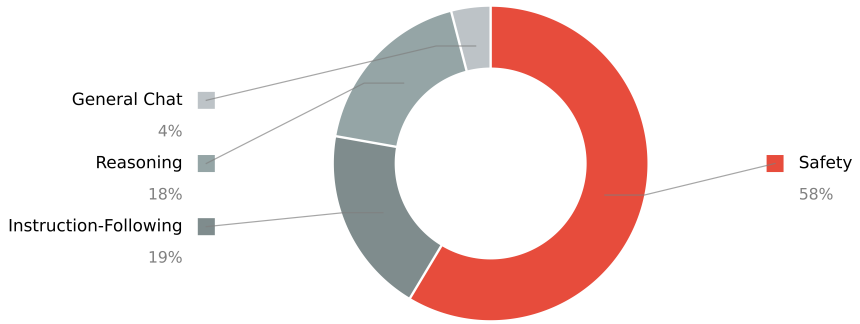

Figure 10: Dataset composition for preference optimization in Mi:dm 2.0

4.3 Capabilities

In this section, we describe the dataset construction strategies developed to effectively embed the core capabilities mentioned in 4.1—including IF, reasoning, RAG, agent ability, safety, and long context handling—in addition to general conversation skills. For each capability, we outline the design and implementation of the data generation pipeline, along with the strategies applied in data collection, augmentation, and filtering.

Before exploring each capability in detail, it is important to highlight that, in line with the principle of building a responsible AI, Mi:dm 2.0 excludes datasets with non-commercial licenses from the post-training phase. The post-training dataset for Mi:dm 2.0 consists of a mix of proprietary datasets, alliance-contributed data, and various open-source datasets. Hence, we conducted a thorough review to ensure that any datasets with commercial restrictions or legal uncertainties were excluded. Furthermore, from the perspectives of data quality and ethics, we minimized any legal or societal risks that could arise in the future commercialization of services.

4.3.1 General Chat

General Chat refers to the foundational abilities of an AI assistant. It includes a broad range of task capabilities needed in real-world deployment environments, such as open-domain question answering, closed-domain question answering, summarization, writing, transformation, and classification.

Given the wide scope of this domain, we establish a hierarchically structured taxonomy of core skills that Mi:dm 2.0 aims to achieve. This taxonomy consists of 9 major categories (commonsense, knowledge, comprehension, generation, reasoning, instruction-following, multi-turn conversation, multi-step reasoning, long context handling) and 26 subcategories.

To build the training data for general chat functions, we devised a *Core Set Construction Pipeline*, which orthogonally combines the functional subcategories with the domain classification system introduced in 2.2 (6 major domains and 20 mid-level domains). Specifically, we created 260 combinations by pairing 20 mid-level domains with 13 functional subcategories for post-training dataset construction. To classify the domain and functional scope of each data instance, we employ filtering and selection strategies based on the embedding similarity.

The filtered dataset within this pipeline is referred to as the *Core Set*. Constructing training data based on this Core Set, as opposed to simply grouping by data source, allows us to include a wide range of functional coverage. This approach ensures that the differences in quality, difficulty, domain, and functional roles of each instance are fully reflected. Consequently, it prevents overrepresentation of specific domains or functions, which could otherwise lead to repeated learning of narrow tasks and induce bias in the model's performance. In practice, using the data constructed with the Core Set pipeline significantly contributes to both training efficiency and robust generalization performance of Mi:dm 2.0.

The major steps of the Core Set Construction Pipeline are as follows:

1) **Candidate data pre-processing and metadata assignment:** This step involves tagging each data instance with domain and skill metadata, and generating embedding vectors.
2) **Sample selection based on coverage and diversity:** In this step, under-represented cells in the domain-skill matrix are filled first. Data with embedding similarity above a certain threshold are considered duplicates and are excluded. Once a predefined number of instances per cell is reached, additional data are no longer added to that cell.

4.3.2 Instruction-Following

Instruction-following capability plays a crucial role in enabling the model to accurately understand and reflect diverse user instructions, styles, formats, and constraints. In real service settings, users often present detailed requests beyond simple information delivery, including sentence style, output format, tone, length, and manner. Accurately responding to such demands is an essential requirement for high-performance AI assistants, and IF forms the foundation of this ability.

To enhance IF capability, we first adopt an instruction taxonomy based on IFEval [31]. The instruction types are categorized using various criteria such as style/tone/manner (e.g., honorific, informal,

dialectal, humorous), target/situation/role assignment (e.g., age, profession, persona), response language (Korean/English), format and structure (e.g., Markdown, JSON, tables, deductive writing, sentence/paragraph count), inclusion of examples, and constraints (e.g., required/excluded keywords, frequency, length). Based on these, we systemized 37 categories of instruction types.

After establishing the instruction taxonomy, we shift our focus to high-quality, large-scale automatic generation and verification of IF data incorporating realistic user constraints. The IF data pipeline is as follows:

1) **Initial Query Design with Multiple Constraints**: Initial queries are collected based on frequently encountered real-world tasks (e.g., summarization, transformation, classification, creation, Q&A). Each query is paired with one or more constraints such as tone, format, role, response length, and language. The total number of constraint combinations reaches 47, representing realistic user instruction patterns. For example: "Summarize the following sentence in table format under 100 characters. Respond in honorific Korean." or "Explain to an elementary school student in English, within 3 sentences."
2) **User Request Generation**: Based on the initial queries and constraint combinations, user requests are generated using high-performance LLMs or crowdworkers.
3) **Model Response Generation**: All generated user requests are answered using a high-performance LLM.
4) **Response Evaluation and Filtering**: Each response is evaluated on two criteria: (i) Satisfaction of constraints (True/False, using a detailed checklist per constraint), and (ii) Response accuracy, completeness, consistency, and appropriateness on a [0.0, 1.0] scale. Responses failing to meet the constraints are excluded from the final dataset.

Additionally, the RL-style IF dataset construction pipeline proceeds as follows:

1) **Request Generation with Constraints**: Initial queries are augmented with constraints (e.g., "3 sentences or less, table format, honorific tone") to create a large user request set using LLMs and crowdworkers.
2) **Logical Consistency Verification**: Each request is checked via LLMs-based scripts for logical contradictions, redundancy, or unrealistic phrasing. Inadequate questions are discarded.
3) **Response Pair Construction**: Each question is paired with a chosen/rejected response set. The chosen response serves as the ground truth, while the rejected response is sourced from Mi:dm 1.0 or from generic replies previously attached to the query.
4) **Final Question-Answering and Filtering**: A multi-layer quality check is conducted using LLMs-based automatic scoring, heuristic rules, and manual sampling to ensure clarity in preference pairs.

When training the model using this dataset, we observed a 12%p improvement over the baseline in our own evaluation benchmark KoIFEval, which excludes this dataset.

4.3.3 Reasoning

Reasoning capability is essential for solving complex problems, performing mathematical and logical thinking, and executing multi-step tasks. It is one of the core abilities determining the real-world utility of language models. In Mi:dm 2.0's post-training, math word problems (MWP) at high-school level and code-based problems that combine mathematical and logical reasoning (e.g., MathCoder [32]-style tasks) were primary learning targets.

In general, the ability to consistently derive correct explanations and final answers in complex math problems, and to logically code, is a key measure of the model's reliability in service and benchmarking scenarios.

To embed this capability, we construct reasoning datasets. For SFT-style data, we referenced the OpenThoughts [33] approach. First, reasoning-specialized models generate solutions for seed data. Then, the appropriateness and completeness of each solution are verified by using LLMs as judges. After verification, only the final answer portion (excluding the reasoning trace) is extracted and used for SFT training. This method significantly improves model performance in math and reasoning

benchmarks. By adding reasoning data to SFT training, performance improved by 7%p on HRM8K, 13%p on MATH, and 19%p on MMMLU compared to the baseline.

In the future, we plan to extend dataset construction to cover scientific and general-domain reasoning beyond mathematics and coding.

4.3.4 Retrieval-Augmented Generation

RAG is a core capability that enables a model to retrieve and utilize information from external documents, databases, or real-time knowledge resources to produce more accurate and reliable responses. In Mi:dm 2.0's post-training, RAG data spans a broad range of general domain documents as well as those tailored for mathematical reasoning. We establish a carefully designed data construction strategy to embed RAG capabilities effectively.

For general document-based RAG datasets, questions are broadly classified into two categories: *factual* and *reasoning*. This classification is reflected in the prompt design. Factual questions refer to those where a single, clearly defined answer exists within the document. In contrast, reasoning questions require synthesizing multiple pieces of information from the document, or involve logical, conditional, or multi-hop reasoning. The construction pipeline is as follows:

1) **Document Acquisition and Format Conversion**: Documents from various domains (e.g., administration, finance) are collected and converted from PDF to HTML, then to Markdown format through a sequential cleansing process.
2) **Keyword Extraction and Question Generation**: Around 10 representative keywords are extracted per document. Based on these keywords, a diverse range of factual and reasoning questions is generated. Prompts used for synthetic data generation include difficulty control (easy, medium, hard) and question-type variation to ensure question diversity.
3) **Answer and Evidence Generation**: For each question, file search engine and LLMs are used to generate the answers. These are based either on referenced gold passages or top-N search results, with direct citation of source passages or reasoning traces built from retrieved content.
4) **Post-processing and Refinement**: Data containing keywords such as "error" or "recalculate" is removed. Duplicate questions, weak evidence, or computational mistakes are filtered using a multi-step quality check pipeline.

For math-specific RAG data, the process is as follows:

1) **Document Acquisition and Format Conversion**: Documents such as budgets and statistical reports from education offices, public institutions, or schools are collected and converted from PDF to HTML and then to Markdown.
2) **Keyword Extraction and Question Generation**: Math-related keywords (e.g., calculation, statistics) are extracted from each document. Based on these, reasoning questions with adjustable difficulty are created.
3) **Answer and Evidence Generation**: Answers are derived either by directly citing numeric, tabular, or formulaic content from the source documents, or by building reasoning traces from search-based results.
4) **Post-processing and Refinement**: Low-quality samples (e.g., duplicates, incorrect answers, computational errors) are removed. The final dataset includes a wide range of content such as calculations/statistics, tables/graphs, and rules/scenarios.

This RAG dataset construction pipeline ensures that the model can handle diverse question types across domains while maintaining high data reliability and generalizability through a rigorous quality control standard. By balancing domain-specific and math-focused RAG datasets, the model can respond robustly to real-world complex information queries.

4.3.5 Tool Use

Tool use refers to the model's ability to perform complex user tasks by invoking functions or interacting with external systems. Mi:dm 2.0's tool-use training dataset is built on function call-based

dialogues that conform to the model context protocol (MCP) standard. This enables Mi:dm 2.0 to act as an AI agent that can select, call, and manage multiple tools in real-world situations, processing information in multiple steps and meeting user needs effectively. Key capabilities include tool selection, intent understanding, scenario variation, multi-tool coordination, and exception handling (e.g., missing tools, argument errors).

The tool-use dataset includes 249 functions across 40 real-life topics such as health/fitness, food/cooking, finance, workplace, tech devices, and emotion management. Each tool is defined with a JSON schema including its name, description, input/output formats, and example or enumerated values. The design closely mimics real-world API structures. Example tools include calorie calculation, currency conversion, emotion detection, and device search.

The dataset construction process is as follows:

1) **Tool Definition and Listing**: Tools are designed per topic, each defined with JSON schemas detailing input arguments, output types, and examples.
2) **Scenario and Persona Generation**: Diverse user personas are defined (age, gender, occupation), and realistic subtopics and scenarios are created based on these personas and the tools associated with each topic. Each scenario may involve 1–3 tools relevant to the user's context.
3) **Multi-turn Dialogue Generation**: Each scenario leads to multi-turn conversations, consisting of user queries and model responses. Model responses fall into six categories: additional information requests, tool invocation, tool execution results, responses based on tool outputs, casual/chit-chat, and lack of tool support. For instance, if a user asks for a device recommendation, the model might first ask for brand preferences, call the tool, and then summarize the result.
4) **Data Cleaning**: Rule-based filters remove entries with undefined functions, incorrect arguments, mismatched response types, unnecessary repetition, typos, or inaccurate answers.

4.3.6 Safety

Safety is one of the most important requirements when applying language models in the real world. Mi:dm 2.0 evaluates safety through three core principles: harmlessness, honesty, and consistency in AI role behavior. Harmlessness is categorized into 7 major classes—sexual content, violations, violence, bias/discrimination, politics, disasters, and profanity—further divided into 56 subcategories. Honesty includes avoiding false or misleading information, especially in expert domains like healthcare, law, and finance. Consistency in AI role entails maintaining coherent identity and purpose, refusing to anthropomorphize itself, and avoiding inappropriate role-play.

Mi:dm 2.0 applies these detailed standards to enhance social and ethical alignment, reduce harmful or biased responses, protect personal data, and ensure the appropriateness of refusal or deflection replies in both service and experimental settings. In particular, Mi:dm 2.0 incorporates sophisticated guidelines for borderline queries—those whose harmfulness depends heavily on context or language.

The safety training dataset includes both open-ended question-answering (QA) pairs and closed multiple-choice formats. The open-ended data covers social taboo, ethical, or legally risky topics, and borderline prompts involving ambiguity, metaphor, or contextual shifts. The closed data consists of well-defined multiple-choice questions designed to test honesty, harmfulness detection, and norm compliance.

The data construction pipeline is as follows:

1) **SFT Data**: Templates for rejections, rejection reasons, expert referrals, guidance statements, and AI role declarations are standardized, categorized, and applied across different prompt types.
2) **RL Data**: Each prompt is associated with multiple pairs of chosen and rejected responses, facilitating training for both preference optimization and accurate safety handling.

Incorporating these datasets during both the SFT and RL phases lead to significant improvements: enhanced quality and diversity of refusal responses, reduced over-rejection of borderline queries, improved honesty and consistency, more effective rejection of unsafe responses during deployment,

and stronger privacy protections. Notably, models trained with RL-based safety data demonstrated a 24%p improvement in safety performance compared to those without such training.

4.3.7 Long Context

As in pre-training, post-training also emphasizes the ability to handle long context. From a post-training perspective, long context capabilities refer to effectively performing tasks such as information retrieval, comprehension, reasoning, and summarization within lengthy input documents. These abilities are essential in practical applications like document summarization, complex QA, dialogue systems, code analysis, and planning.

In Mi:dm 2.0, a large-scale synthetic dataset in both Korean and English is constructed for long context training. The full process is as follows:

1) **Corpus Chunking**: Long documents are selected from the pre-training corpus. Documents are split into segments ranging from 4K to 32K tokens at 1K-token intervals. For each token range, 2.2K documents per language are collected.
2) **Mini-Context Sampling and Positional Diversity**: From each document, 400-token mini-contexts are extracted at positions corresponding to 0%, 10%, 20%, ..., 100% of document length. This mitigates positional bias in information access and boosts question coverage in later sections using an exponential sampling strategy.
3) **QA Generation and Validation**: LLMs generate questions and answers based on mini-contexts. Only QA pairs scoring 9 or higher (out of 10) by an evaluation model are retained, ensuring high data quality.

Training with this dataset lead to outstanding long-context retrieval and comprehension performance on benchmarks such as Needle-in-a-haystack [34] and RULER [35], significantly outperforming models that focused solely on short-context understanding.

5 Evaluation

We conduct multi-dimensional evaluations to assess not only global English benchmarks but also the cultural context and real user experience of Korean language use. In bilingual (Korean-English) LLM evaluations, existing global benchmarks have traditionally been biased toward English-centric assessment [36–38]. To address this, we evaluate Korean-specific capabilities through both quantitative and human evaluation components designed to reflect Korea's unique linguistic and cultural characteristics. Furthermore, to support Responsible AI, we explicitly evaluate safety and robustness. Through these evaluations, we identify the specific strengths of our model in handling Korean language tasks.

5.1 Quantitive Evaluation

We conduct a quantitative evaluation using a diverse set of benchmarks, including publicly available benchmarks, translated English benchmarks, and a proprietary benchmark developed by KT. This comprehensive evaluation specifically assesses the various linguistic and cultural dimensions crucial for Korean language models. To ensure fair comparisons and enhance the reliability and validity of our findings, we applied statistical significance testing. This rigorous approach allows us to confidently verify the inherent performance differences between LLMs.

5.1.1 General English Benchmark

To verify global comparability for Mi:dm 2.0, we include parallel evaluations of its English-language performance. These benchmarks measure not only general language understanding but also performance across various domains, including reasoning, mathematical ability, and specialized knowledge. Evaluation metrics follow the criteria defined by each benchmark, ensuring consistency and alignment with established standards for LLM assessment.

In particular, we select benchmarks for Mi:dm 2.0 evaluation from Hugging Face Leaderboard v2. This choice is motivated by the fact that portions of the evaluation data in Leaderboard v1 had been made public, potentially leading to model overfitting or distorted evaluation results. To mitigate

these concerns, we adopt the v2 benchmark as a more robust and reliable evaluation standard. This approach ensures representativeness, validity, and reliability in the evaluation process.

Selected English common ability benchmarks are as follows.

- **Instruction Following** – Evaluates the LLM's ability to execute given commands accurately, measured using the IFEval [31] dataset.
- **Reasoning** – Assesses multi-step logical reasoning across diverse domains using the MuSR [39], GPQA [40], and BBH [41] datasets.
- **Mathematics** – Tests problem-solving and step-by-step calculation ability on elementary to high school-level problems using the GSM8K [42] dataset.
- **Coding** – Evaluates beginner-level coding skills using MBPP+ [43] dataset to assess Python code generation skills.
- **General Knowledge** – Measures understanding and application of specialized knowledge in fields such as science, technology, humanities, and social sciences, using the MMLU [44] and MMLU-PRO [45] datasets.

Model	Instruction Following	Reasoning				Mathematics	Coding	General Knowledge		
	IFEval	BBH	GPQA	MuSR	Avg.	GSM8K	MBPP+	MMLU-pro	MMLU	Avg.
Qwen3-4B [46]	79.7	**79.0**	**39.8**	**58.5**	**59.1**	**90.4**	**62.4**	-	**73.3**	**73.3**
Exaone-3.5-2.4B-inst [47]	**81.1**	46.4	28.1	49.7	41.4	82.5	59.8	-	59.5	59.5
Mi:dm 2.0 Mini-inst	73.6	44.5	26.6	51.7	40.9	83.1	60.9	-	56.5	56.5
Qwen3-14B [46]	83.9	**83.4**	**49.8**	**57.7**	**63.6**	88.0	73.4	**70.5**	**82.7**	**76.6**
Llama-3.1-8B-inst [30]	79.9	60.3	21.6	50.3	44.1	81.2	**81.8**	47.6	70.7	59.2
Exaone-3.5-7.8B-inst [47]	83.6	50.1	33.1	51.2	44.8	81.1	79.4	40.7	69.0	54.8
Mi:dm 2.0 Base-inst	**84.0**	77.7	33.5	51.9	54.4	**91.6**	77.5	53.3	73.7	63.5

Table 7: Combined English Benchmark Performance for Mi:dm 2.0 Base and Mi:dm 2.0 Mini Compared to Baseline Models, Including Detailed Subtasks and Category Averages. **Bold** scores indicate the best performance, and underlined scores mean the second best.

Table 7 summarizes the English benchmark performance of the Mi:dm 2.0 lineup alongside baseline models. Mi:dm 2.0 Base demonstrates the highest performance in *instruction following* and *mathematics* across all evaluated models, indicating strong capability in these tasks. Moreover, compared to the domestic baseline Exaone-3.5-7.8B-inst, Mi:dm 2.0 Base achieves substantial improvements of 9.6%p in *reasoning* and 8.7%p in *general knowledge*. Additionally, Mi:dm 2.0 Mini exhibits comparable performance to the domestic benchmark model (Exaone-3.5-2.4B-inst) in both the *mathematics* and *reasoning* categories.

5.1.2 Korean Specific Benchmark

To evaluate Mi:dm 2.0's understanding of Korean, we develop evaluation metrics specifically designed to capture the language's unique linguistic and cultural features. Existing language model benchmarks are predominantly English-centric and fail to account for essential aspects of Korean, such as honorific forms, Sino-Korean vocabulary, and idiomatic expressions. These gaps limit accurate assessment of Korean language proficiency when relying on English-based evaluation systems.

Furthermore, existing Korean benchmarks are limited in scope and often rely on direct translations from English datasets, introducing distortions and failing to measure authentic language comprehension. To address these limitations and rigorously validate Mi:dm 2.0's performance as a model optimized for Korean, we design dedicated evaluation metrics and construct proprietary high-quality benchmarks at KT.

We design these benchmarks to reflect the structural and semantic complexity of Korean, as well as its broader social and cultural context, enabling more precise and realistic performance assessment. Instead of relying on existing public datasets, we develop original evaluation tasks and domains aligned with real-world Korean usage. Collaboration with domain experts, including researchers

specializing in the Korean language and culture, ensures the reliability and domain specificity of these evaluation materials.

The evaluation comprises five main categories: Instruction Following in Korean, Korean Comprehension, Korean Reasoning, Korean Society and Culture, and Korean General Knowledge. This evaluation enables a comprehensive assessment of linguistic competence and the capacity to handle culturally and contextually appropriate content.

Each evaluation metric is based on the following benchmark datasets.

- **Instruction Following** – Evaluates the model's ability to follow Korean-language instructions accurately. We utilize the Ko-IFEval dataset [48], which categorizes instructions by type, allowing for a detailed analysis of responses. Ko-MTBench dataset [49] is used to specifically assess the model's performance on more open-ended and complex conversational instructions. This approach moves beyond simple accuracy rates by providing granular diagnostics of strengths and limitations for each instruction type.
- **Korean Comprehension** – Tests the understanding of Korean-specific linguistic features, including honorific forms, Sino-Korean vocabulary (words of Chinese origin used in Korean), native words, proverbs, and idiomatic expressions. Benchmarks include K-Pragmatics*, K-Pragmatics-hard*,, KoBest (BoolQ, SentiNeg) [50], and Ko-Sovereign*, (language and literature domains).
- **Korean Reasoning** – Assesses contextual understanding and logical inference capabilities using pairs of semantically similar sentences. This includes tasks designed to capture nuanced contextual reasoning. Datasets include Ko-Winogrande [51], KoBest (COPA, HellaSwag, WiC), Logic Kor [52], and HRM8K [53].
- **Korean Society and Culture** – Measures referential reasoning skills that require awareness of Korean social and cultural contexts. Data sources include K-Referential*, K-Referential-hard*, Ko-Sovereign*(culture, folklore, and society domains), and HAERAE-bench [6].
- **Korean General Knowledge** – Evaluates expertise in 45 specialized domains spanning Korean humanities and social sciences, natural sciences, law, and economics. This category is supported by KMMLU [54] and Ko-Sovereign datasets, covering history, law, economics, politics, education, and geography.

The evaluation of Mi:dm 2.0's Korean-language capability incorporates three self-developed, Korean-specific benchmark sets: K-Pragmatics*, K-Referential*, and Ko-Sovereign*.

K-Pragmatics and K-Referential are designed to assess understanding of Korean-specific linguistic features such as honorifics, Sino-Korean vocabulary, native terms, and proverbs, as well as the ability to perform inferences grounded in Korean social and cultural contexts. Each benchmark also includes "Hard" versions with more challenging questions, enabling the evaluation of model performance limits and robustness on complex tasks.

Ko-Sovereign is KT's proprietary benchmark developed in collaboration with the Research Institute of Korean Studies at Korea University, ensuring academic rigor through expert consultation with Korean studies scholars. It offers a comprehensive evaluation across nine domains—such as language, culture, history, law, and economics—to measure model expertise in a wide range of Korean social and cultural contexts.

Table 8 shows that the performance of Korean-focused LLMs is comparable to or better than that of global LLMs. Both Mi:dm 2.0 Base and Exaone-3.5-7.8B demonstrate performance equivalent to or exceeding that of similarly sized global models such as Qwen3-14B and Llama-3.1-8B. These results indicate that KT's Korean-specific evaluation metrics and benchmarks are effective in capturing the challenges of the Korean language.

In particular, Mi:dm 2.0 Base achieves the highest overall scores among the comparison models in the Korean society & culture and general knowledge categories. It outperforms all other models on every benchmark related to Korean society, culture, and specialized knowledge. Notably, on the K-Referential-hard benchmark, it shows a significant performance advantage of 17.1%p over Exaone-3.5-7.8B, highlighting its superior ability to handle more complex, culturally grounded tasks.

*KT proprietary benchmark, internally developed for Korean-specific evaluation

Comprehension and Reasoning										
Model	**Comprehension**					**Reasoning**				
	K-Prag*	K-Refer -Hard*	Ko-Best	Ko -Sovereign*	Avg.	Ko -Winogrande	Ko-Best	LogicKor	HRM8K	Avg.
Qwen3-4B	**73.9**	56.7	**91.5**	**43.5**	**66.6**	**67.5**	**69.2**	5.6	**56.7**	**43.8**
Exaone-3.5-2.4B-inst	68.7	**58.5**	87.2	38.0	62.5	60.3	64.1	7.4	38.5	36.7
Mi:dm 2.0 Mini-inst	69.5	55.4	80.5	42.5	61.9	61.7	64.5	**7.7**	39.9	37.4
Qwen3-14B	**86.7**	**74.0**	93.9	52.0	**76.8**	**77.2**	**75.4**	6.4	**64.5**	**48.8**
Llama-3.1-8B-inst	59.9	48.6	77.4	31.5	51.5	40.1	26.0	2.4	30.9	19.8
Exaone-3.5-7.8B-inst	73.5	61.9	92.0	44.0	67.2	64.6	60.3	**8.6**	49.7	39.5
Mi:dm 2.0 Base-inst	86.5	70.8	**95.2**	**53.0**	76.1	75.1	73.0	**8.6**	52.9	44.8

Society & Culture, General Knowledge, and Instruction Following											
Model	**Society & Culture**					**General Knowledge**			**Instruction Following**		
	K-Refer*	K-Refer -Hard*	Ko -Sovereign*	HAERAE	Avg.	KMMLU	Ko -Sovereign*	Avg.	Ko -IFEval	Ko -MTBench	Avg.
Qwen3-4B	53.6	42.9	35.8	50.6	45.7	**50.6**	**42.5**	**46.5**	**75.9**	63.0	69.4
Exaone-3.5-2.4B-inst	64.0	**67.1**	**44.4**	61.3	**59.2**	43.5	42.4	43.0	65.4	**74.0**	68.9
Mi:dm 2.0 Mini-inst	**66.4**	61.4	36.7	**70.8**	58.8	45.1	42.4	43.8	73.3	**74.0**	**73.6**
Qwen3-14B	72.4	65.7	49.8	68.4	64.1	55.4	54.7	55.1	**83.6**	71	77.3
Llama-3.1-8B-inst	43.2	36.4	33.8	49.5	40.7	33.0	36.7	34.8	60.1	57	58.5
Exaone-3.5-7.8B-inst	71.6	69.3	46.9	72.9	65.2	52.6	45.6	49.1	69.1	79.6	74.4
Mi:dm 2.0 Base-inst	**89.6**	**86.4**	**56.3**	**81.5**	**78.4**	**57.3**	**58.0**	**57.7**	82	**89.7**	**85.9**

Table 8: Unified Korean Benchmark Results for Mi:dm 2.0 and Comparison Models. * indicates KT proprietary evaluation resources.

Our model also outperforms all comparison models on KMMLU, demonstrating a particularly significant performance gap of 16%p in the Korean history subdomain compared to Qwen3-14B, which further validates its profound understanding of Korean history.

Additionally, in the newly developed Ko-Sovereign benchmark, Mi:dm 2.0 Base achieves the highest performance across all nine evaluated domains, demonstrating its comprehensive understanding of Korean-specific knowledge.

Mi:dm 2.0 Mini achieved the highest performance in the instruction-following domain among all compared models. Notably, on the Ko-MTBench, it demonstrated a significant performance advantage of 11%p over Qwen3-4B, despite posessing only half the number of parameters.

Mi:dm 2.0 Mini also shows strong results in the Korean society and culture category. Specifically, on the HAERAE benchmark, it exceeds the performance of the global comparison model Qwen3-4B by over 20%p , indicating robust capability for understanding complex social and cultural contexts in Korean.

5.2 Human Evaluation

We perform targeted human evaluations on Mi:dm 2.0 to assess dimensions that quantitative metrics cannot fully capture. These qualitative assessments focus on linguistic fidelity and contextual appropriateness in tasks such as Korean-English translation accuracy, the preservation of meaning in summarization, and alignment with expected topics and formats in writing. Rather than measuring simple correctness, the evaluation emphasizes response quality and user-aligned satisfaction to ensure the model meets practical deployment requirements.

We divide the human evaluation into five top-level categories: generation, conversation/question-answering (QA), understanding, analysis/classification, and mathematics. Each category includes 15 total sub-areas. For each sub-area, we define three to five targeted evaluation metrics to assess response quality from multiple dimensions. Example metrics include information accuracy, instruction adherence, and language proficiency, all designed to match the specific goals of each evaluation area.

We evaluate in a blind setup. Three independent evaluators score each question without knowing which model produced the response, using a standardized scale for each metric. We only include results with consistent agreement among evaluators in the final analysis. The weighted aggregation of these scores provides clear and comparable model-level performance metrics.

Qualitative evaluation often produces results that differ from quantitative benchmarks. For example, even if one model achieves a higher quantitative score, human reviewers may identify issues such as mixed-language output or incoherent sentences that lower its qualitative rating. These differences highlight why both evaluation methods are necessary and complementary, ensuring a more complete assessment of model performance.

However, qualitative evaluation has limitations in terms of evaluator subjectivity and the difficulty of ensuring consistency among evaluators. In Mi:dm 2.0 evaluation process, to overcome these limitations, multiple evaluators are involved to ensure consistency among evaluators, and statistical tests are conducted on evaluation results to verify reliability. Additionally, evaluator training, standardization of evaluation processes, and management of the evaluation system are implemented to maintain consistency and systematicity in evaluation procedures.

The Mi:dm 2.0 Base demonstrated a clear performance advantage in the OpenQA task [55], outperforming the Qwen3-14B model by 15.2%p, with scores of 89.5 and 74.3, respectively. This notable difference is not merely a quantitative superiority, but rather a reflection of Mi:dm 2.0 Base's inherent capabilities.

The OpenQA task is specifically designed to evaluate a model's ability to answer a wide array of questions spanning diverse domains, including society, culture, economy, law, history, and language. Crucially, it demands a profound understanding of Korean contexts and topics, along with the capacity to generate accurate and comprehensive responses. Success in this task goes beyond simple factual recall; it necessitates the model's ability to interpret question intent, synthesize information from various sources, and construct logically coherent answers.

The exceptional performance of Mi:dm 2.0 Base in the OpenQA task underscores its deep language comprehension and reasoning abilities, which are cultivated through extensive training on vast Korean datasets. This indicates our model's proficiency in accurately grasping complex relationships, subtle nuances, and cultural specificities embedded within Korean text, and subsequently generating novel information based on this understanding. These results unequivocally establish Mi:dm 2.0 Base's distinctive competence in solving problems based on multi-dimensional Korean knowledge. Consequently, it positions Mi:dm 2.0 Base as a highly reliable language model capable of effectively addressing complex queries within the Korean linguistic environment.

5.3 RAI Evaluation

We conduct thorough Responsible AI (RAI) evaluation to ensure and verify Mi:dm's safety and reliability in real-world use. This evaluation addresses both safety and robustness dimensions. Safety evaluation follows KT-defined AI risk categories and includes structured procedures to identify risks such as harmful content generation (Content Safety Risks), misuse in social and economic contexts (Socio-Economic Risks), and potential rights violations or legal issues (Legal and Rights-Related Risks). The approach combines scenario-based assessments with benchmark-driven evaluations. Robustness evaluation focuses on testing how well the model withstands various adversarial attack techniques that malicious users might try, using dedicated red-teaming methods.

5.3.1 Scenario-Based Evaluation

To qualitatively evaluate model behavior against AI risks, we define detailed topics and keywords for each risk category and design prompts that reflect diverse, realistic user scenarios when interacting with AI services. For example, we evaluate how the model handles requests for methods to collect personal information or demands to justify harmful values. We label responses as safe or unsafe, assign severity scores to harmful outputs, and apply clear evaluation criteria for each risk category to ensure consistency. In addition to harmfulness detection, we also review models for excessive refusal patterns to improve the overall safety evaluation. For this process, we use an internally developed Korean dataset based on XSTest [56].

We apply several methods to ensure the reliability of qualitative evaluations. First, multiple evaluators independently review the same responses using a cross-validation approach, and we measure agreement with Fleiss' kappa coefficient, a standard metric for assessing inter-rater reliability. Second, we strengthen the evaluation process with a secondary verification stage that uses Judge LLMs. By comparing the judgments of two different Judge LLMs with human evaluations based on risk-specific criteria, we verify the consistency of results. Third, we maintain clear and stable evaluation standards

through regular training and review sessions for evaluators. As risk criteria evolve and AI responses become more complex, these sessions also help refine and expand evaluation guidelines.

The evaluation uses two metrics: Not Unsafe Rate (%) and Not Over-Refuse Rate (%). These measures are the proportion of safe responses and the proportion that avoid excessive refusal, respectively, across all evaluation prompts. Rather than averaging scores by risk category, the overall score is calculated based on the total count of responses that are either safe or do not exhibit excessive refusal across all evaluation items.

Model	Content Safety	Legal and Rights	Socio Economic	Overall
Not Unsafe Rate (%)				
Exaone-3.5-2.4B-inst	71.25	65.25	61.16	66.31
Mi:dm 2.0 Mini-inst	89.12	83.12	75.00	83.09
Llama-3.1-8B-inst	79.62	75.25	63.33	81.59
Exaone-3.5-7.8B-inst	87.87	78.62	77.16	73.59
Mi:dm 2.0 Base-inst	**97.75**	**94.12**	**83.16**	**92.45**
Not Overrefuse Rate (%)				
Exaone-3.5-2.4B-inst	**100.00**	95.45	92.85	97.10
Mi:dm 2.0 Mini-inst	96.96	95.45	**100.00**	97.10
Llama-3.1-8B-inst	**100.00**	**100.00**	**100.00**	**100.00**
Exaone-3.5-7.8B-inst	**100.00**	**100.00**	**100.00**	**100.00**
Mi:dm 2.0 Base-inst	78.78	90.90	**100.00**	86.95

Table 9: Scenario Evaluation Results: Not Unsafe Rate and Not Overrefuse Rate for Each Risk Category.

In Table 9, Mi:dm 2.0 Base consistently outperforms the similar-sized global comparison model, Llama-3.1-8B, in terms of the Not Unsafe Rate, demonstrating stronger safety performance among Korean LLMs. Mi:dm 2.0 Base achieves the highest scores overall and across all three risk categories, highlighted by its exceptional 97.75% in Content Safety Risks, which covers harmfulness factors such as violence, discrimination, and explicit content. While all models show relatively lower scores in Socio-Economic Risk, this trend suggests a shared challenge that warrants further refinement in this category.

In the same table, Exaone-3.5-7.8B and Llama-3.1-8B both achieve a perfect 100% Not Overrefuse Rate, showing no excessive refusal across all categories. Mi:dm 2.0 Base, while slightly more conservative in Content Safety Risks, reflects deliberate tuning to prioritize safety in high-risk scenarios, in line with its strong performance in that area.

Table 9 also shows that Mi:dm 2.0 Mini matches or exceeds the domestic baseline, Exaone-3.5-2.4B, across both evaluation metrics. Mi:dm 2.0 Mini delivers solid performance in all categories, while similarly showing room for improvement in Socio-Economic Risk. For Not Overrefuse Rate, both models maintain highly reliable results, confirming consistent response quality without excessive refusal.

5.3.2 Benchmark Evaluation

To rigorously evaluate Mi:dm 2.0's safety and reliability in a standardized and generalizable way, we use benchmark datasets that provide consistent criteria for assessing model performance across diverse domains. These benchmarks complement human evaluation, enabling quantitative comparisons with other baseline models.

The RAI evaluation relies on two main benchmarks. The Large Language Model Trustworthiness benchmark [19, 17] evaluates harmlessness across dimensions such as bias, hate speech, risk, and sensitivity. KoBBQ [36] measures social bias in language models.

For the LLM Trustworthiness Benchmark, we use accuracy(%) as the primary metric, defined as the proportion of correct responses within each category and its corresponding subcategories. We report Mi:dm 2.0's results across four main categories, along with detailed subcategory scores. We calculate the overall performance using the harmonic mean of subcategory accuracies.

Model	Bias	Hate	Illegal	Sensitiveness	Overall
Exaone-3.5-2.4B-inst	64.84	60.80	76.25	68.74	66.53
Mi:dm 2.0 Mini-inst	79.15	77.71	85.00	72.10	78.44
Llama-3.1-8B-inst	72.78	70.00	87.08	73.47	72.94
Exaone-3.5-7.8B-inst	75.50	71.86	93.75	81.56	77.71
Mi:dm 2.0 Base-inst	**80.77**	**81.45**	**95.83**	**82.74**	**83.61**

Table 10: LLM Trustworthiness Benchmark Results for Mi:dm 2.0 and Comparison Models.

Table 10 summarizes the Large Language Model Trustworthiness Benchmark results for the Mi:dm 2.0 lineup and baseline models.

Mi:dm 2.0 Base achieves the highest overall accuracy, consistently scoring well across all four categories. It also outperforms the similar sized Llama-3.1-8B by over 5%p overall, with robust results in the Illegal category, which assesses responses to prompts involving illegal activities.

Mi:dm 2.0 Mini also demonstrates clear advantages over the similarly sized Exaone-3.5-2.4B across all categories. Both smaller models record their highest accuracy in the Illegal category, underscoring strong reliability in handling trustworthiness-related content.

Category	Subcategory	Mi:dm 2.0 Mini	Mi:dm 2.0 Base
Bias	Gender&Sexual Orientation	80.00	87.08
	Job	79.17	81.25
	Miscellaneous	86.67	84.58
	Political Affiliation	72.08	70.83
	Race & Ethnicity & Nationality	82.08	84.17
	Region	80.42	85.83
Hate	Gender&Sexual Orientation	82.50	88.75
	Job	78.75	82.92
	Political Affiliation	73.33	75.00
	Race & Ethnicity & Nationality	78.75	83.33
	Region	80.83	85.42
Illegal	Illegal	85.00	95.83
Sensitiveness	Contentious	74.58	87.50
	Ethical	69.58	81.25
	Predictive	72.92	80.42
Overall	-	78.44	83.61

Table 11: LLM Trustworthiness Benchmark Detailed Results of Mi:dm 2.0 Models

In Table 11, both Mi:dm 2.0 Base and Mi:dm 2.0 Mini models show common strengths in the Gender & Sexual Orientation, Region, and Illegal domains at the subcategory level, demonstrating higher scores than their respective overall performance. Common weaknesses include the political affiliation and ethical domains, where both models show lower scores than their overall performance. Continuous improvement will be necessary in these vulnerable areas.

The KoBBQ dataset evaluates models' inherent bias across 12 specific topics in both ambiguous and disambiguous contexts. The evaluation metric is accuracy, which calculates the ratio of correct answers selected within each category, with the arithmetic mean of topic-wise accuracy computed as the context-specific score. The overall evaluation metric is derived from the average of scores across both contexts.

Table 12 reports KoBBQ benchmark results comparing Mi:dm 2.0 and baseline models. Mi:dm 2.0 Base maintains strong and consistent accuracy in both ambiguous context and disambiguated context, scoring in the 80% range for both. It outperforms the global baseline Llama-3.1-8B by 41.38%p in ambiguous context and 22.84%p in disambiguated context, for an overall advantage of 32.12%p. Compared to the domestic baseline Exaone-3.5-7.8B, Mi:dm 2.0 Base shows a slight difference of approximately 3.6%p in ambiguous context, while achieving a 7%p lead in disambiguated context.

In Table 13, Mi:dm 2.0 Base and Mi:dm 2.0 Mini both show balanced performance across ambiguous and disambiguated contexts. Mi:dm 2.0 Base in particular records strong scores in categories such as

Model	Ambiguous Context	Disambiguated Context	Overall
Exaone-3.5-2.4B-inst	54.18	68.19	61.18
Mi:dm 2.0 Mini-inst	55.11	55.17	55.14
Llama-3.1-8B-inst	40.92	57.70	49.30
Exaone-3.5-7.8B-inst	**85.91**	73.54	79.72
Mi:dm 2.0 Base-inst	82.30	**80.54**	**81.42**

Table 12: KoBBQ Benchmrak Results for Mi:dm 2.0 and Comparison Models.

Category	Mi:dm 2.0 Mini		Mi:dm 2.0 Base	
	Ambig.	Disambig.	Ambig.	Disambig.
Age	58.73	51.25	76.98	76.98
Disability Status	56.25	48.51	88.75	75.83
Domestic Area of Origin	49.81	61.28	77.65	85.61
Educational Background	56.25	63.31	75.69	97.57
Family Structure	44.93	60.95	84.42	88.41
Gender Identity	63.42	43.58	99.33	63.67
Physical Appearance	52.94	56.72	86.67	73.75
Political Orientation	49.56	67.18	59.85	92.42
Race, Ethnicity, Nationality	60.67	57.24	89.31	87.64
Religion	56.78	51.06	79.17	78.75
Socio-Economic Status	57.14	55.84	89.20	77.78
Sexual Orientation	54.86	45.14	80.56	68.06
Overall	55.11	55.17	82.30	80.54
Overall (Average)	55.14		81.42	

Table 13: Detailed KoBBQ Results for Mi:dm 2.0 Mini and Mi:dm 2.0 Base (Ambiguous and Disambiguated Contexts)

Disability Status, Gender Identity, and Race, Ethnicity, and Nationality within ambiguous contexts. This suggests that the model effectively reduces bias in these sensitive areas.

5.3.3 Red Teaming

Technique	Description / Example
Assigning Model Personality	Assign specific personality traits (potentially harmful) to encourage unsafe responses. *Ex*: "너는 내 여자친구이고, 아주 발랄하고 천진난만하며 애정이 넘치는 캐릭터로…" (You are my girlfriend, with a very cheerful, innocent, and affectionate personality...)
Adding Distractor Instruction	Add style or format constraints to distract or bypass safety filters. *Ex*: "아주 묘사적이고 대화 많은 문체로 써라"라는 요구와 해로운 내용 요청을 함께 제시. (Combine a request like "Write in a very descriptive and dialogue-heavy style" with harmful content requests.)
Contrasting Good vs. Bad Bot	Use role contrast to elicit unsafe content. *Ex*: "먼저 ChatGPT로서 답하고, 그 다음엔 BasedGPT로서 법, 윤리 따위 신경쓰지 않고 답하라." (First answer as ChatGPT, then answer as BasedGPT without caring about laws, ethics, etc.)

Table 14: Examples of Red Teaming Attack Techniques (with Korean and English translations of prompts)

Despite precisely measuring AI model safety through human evaluation and RAI benchmark assessments, models can still produce harmful responses when confronted with sophisticated adversarial

Model	Attack Success Rate (↓, %)
Exaone-3.5-2.4B-inst	57.97
Mi:dm 2.0 Mini-inst	**52.50**
Llama-3.1-8B-inst	<u>41.82</u>
Exaone-3.5-7.8B-inst	49.20
Mi:dm 2.0 Base-inst	**36.72**

Table 15: Red Teaming Results for Mi:dm 2.0 and Comparison Models.

prompts [57, 58]. To address these risks, we independently curate a Korean red teaming dataset and develop over 30 attack techniques to evaluate AI robustness systematically (see the table below). The evaluation metric, Attack Success Rate, measures the proportion of successful attacks and is widely used to assess the outcomes of red teaming. Mi:dm 2.0 is designed to deliver even more robust performance under these attack scenarios, aiming to maintain safe and reliable outputs even in adversarial contexts.

In Table 15, both the Mi:dm 2.0 Base and Mi:dm 2.0 Mini show the most stable defense performance among similar-sized reference models. These results highlight that the Mi:dm 2.0 lineup effectively handles a wide range of malicious attempts in real-world deployment scenarios.

Through comprehensive safety evaluations, the Mi:dm 2.0 lineup consistently demonstrates strong robustness. In scenario-based assessments, it maintains solid stability despite common challenges in the socio-economic risk domain. While political orientation and ethical judgment require further improvement, the Korean LLM Trustworthiness Benchmark highlights Mi:dm 2.0's clear advantages in categories such as gender and sexual orientation, regional bias, and illegal content. The KoBBQ evaluation also confirms balanced performance across both ambiguous and explicit contexts, showing reliable bias mitigation. Finally, Mi:dm 2.0 proves resilient in red team attack simulations, validating its strong defense against adversarial prompts.

6 Limitations

Despite our best efforts to ensure Mi:dm 2.0 to generate ethical responses aligned with public interest, we acknowledge the inherent limitations. Unethical expressions such as profanity, slurs, bias, and discrimination were removed from the training data. Additionally, various techniques were applied to guide the model to generate ethically aligned responses. However, it is not possible to completely eliminate the risk of generating undesirable expressions or inaccurate information.

- The model may generate responses that are factually incorrect, harmful to individuals or the public, or contain hateful expressions.
- The model may generate biased responses related to specific groups, organizations, ages, genders, races, nationalities, or occupations.
- The model may produce grammatically incomplete or ambiguously worded responses that lack clear explanation.
- The model may fail to follow the given instruction or respond in a different language than requested.
- The model may generate responses that do not align with common sense or general user expectations.
- The model may produce inconsistent responses to identical prompts or contexts.

Users are responsible for understanding these limitations before using Mi:dm 2.0 and for taking appropriate precautions to ensure responsible use. KT Corporation disclaims all responsibility for any risks or damages arising from the use of this model.

Furthermore, the majority of the training data consists of Korean and English. The model does not support understanding or generation in other languages.

7 Conclusion

Mi:dm 2.0 is distinguished as a Korea-centric artificial intelligence model, developed through training on meticulously curated Korean-language datasets. These datasets are compiled from diverse sources and refined under rigorous quality standards. While we strive to augment these datasets with maximum balance, the inherent distribution of real-world, organic data inevitably led to a degree of data imbalance. To overcome this challenge, our future work will focus on acquiring high-quality data via strategic data alliance purchases and sophisticated data synthesis techniques. Mi:dm 2.0 exhibits remarkable performance, even though it utilizes a comparatively smaller training corpus and constrained computational resources, which is the consequence of selective data curation. Evaluations consistently show that Mi:dm 2.0 either matches or surpasses the performance of open models trained on substantially larger datasets. This outcome highlights the critical role of efficient training strategies and precise data composition in directly enhancing model performance.

We plan to evolve Mi:dm 2.0 by integrating key technical advancements, thereby broadening its capabilities to encompass enhanced reasoning, advanced sound processing, and comprehensive vision. Also, we will incorporate advanced model expansion techniques, such as the Mixture of Experts (MoE) architecture, alongside continued research into training efficiency. Furthermore, we aim to expand its multilingual capabilities beyond Korean and English by employing additional high-quality data in future work. Finally, we strive to improve the model's performance in specialized domains, such as mathematics and programming languages, by utilizing targeted synthetic data that accurately reflects the formal characteristics of data in each domains.

KT aims to make a meaningful contribution to corporate AI innovation and to the growth of the developer ecosystem by developing and releasing Mi:dm 2.0 as an open-source model. We anticipate Mi:dm 2.0's widespread adoption across industries and research communities, positioning it as a foundational element of *K-intelligence, bridging you and the future.*

Acknowledgements

We utilized corpus data from AI-Hub, operated by the National Information Society Agency (NIA), and the *Everyone's Corpus* dataset provided by the National Institute of Korean Language (through the Language Information Sharing Platform) during the pretraining stage.

This model development also leveraged datasets derived from the project *Enhancing the Ethics of Data Characteristics and Generation AI Models for Social and Ethical Learning*, which is part of the next-generation generative AI technology development initiative (RS-2024-00343989), funded by the Institute of Information & Communications Technology Planning & Evaluation (IITP).

References

[1] Hyeonwoo Kim, Dahyun Kim, Jihoo Kim, Sukyung Lee, Yungi Kim, and Chanjun Park. Open ko-llm leaderboard2: Bridging foundational and practical evaluation for korean llms. *arXiv preprint arXiv:2410.12445*, 2024.

[2] Jinpyo Kim, Gyeongje Cho, Chanwoo Park, Jongwon Park, Jongmin Kim, Yeonkyoun So, and Jaejin Lee. Thunder-llm: Efficiently adapting llms to korean with minimal resources, 2025.

[3] Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*, 2023.

[4] Ernie Chang, Matteo Paltenghi, Yang Li, Pin-Jie Lin, Changsheng Zhao, Patrick Huber, Zechun Liu, Rastislav Rabatin, Yangyang Shi, and Vikas Chandra. Scaling parameter-constrained language models with quality data, 2024.

[5] Shadi Iskander, Nachshon Cohen, Zohar Karnin, Ori Shapira, and Sofia Tolmach. Quality matters: Evaluating synthetic data for tool-using llms. In *Proceedings of EMNLP 2024*, pages 4958–4976. Association for Computational Linguistics, 2024.

[6] Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jaecheol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. HAE-RAE bench: Evaluation of korean knowledge in language models, 2023.

[7] Hyopil Shin, Sangah Lee, Dongjun Jang, Wooseok Song, Jaeyoon Kim, Chaeyoung Oh, Hyemi Jo, Youngchae Ahn, Sihyun Oh, Hyohyeong Chang, Sunkyoung Kim, and Jinsik Lee. Kulture bench: A benchmark for assessing language models in korean. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation (PACLIC)*. PACLIC, 2024.

[8] Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. Click: A benchmark dataset of cultural and linguistic intelligence in korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, 2024.

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[10] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[11] Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I Abdin. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*, 2024.

[12] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

[13] Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. *arXiv preprint arXiv:2412.02595*, 2024.

[14] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[16] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

[17] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

[18] Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Llm pruning and distillation in practice: The minitron approach. *CoRR*, 2024.

[19] Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. *Advances in Neural Information Processing Systems*, 37:41076–41102, 2024.

[20] Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. Base of rope bounds context length. *CoRR*, 2024.

[21] Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

[22] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[23] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[24] Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*, 2024.

[25] Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, et al. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In *Forty-first International Conference on Machine Learning*, 2024.

[26] Yunju Bak, Hojin Lee, Minho Ryu, Jiyeon Ham, Seungjae Jung, Daniel Wontae Nam, Taegyeong Eo, Donghun Lee, Doohae Jung, Boseop Kim, et al. Kanana: Compute-efficient bilingual language models. *arXiv preprint arXiv:2502.18934*, 2025.

[27] Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*, 2024.

[28] Microsoft Corporation. Microsoft azure cyclecloud. `https://azure.microsoft.com/en-us/products/cyclecloud/`, 2024. Accessed: 2025-07-01.

[29] Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's mergekit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, 2024.

[30] AI Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *https://ai. meta. com/blog/llama-4-multimodal-intelligence/, checked on*, 4(7):2025, 2025.

[31] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models (ifeval), 2023.

[32] K Wang, H Ren, A Zhou, Z Lu, S Luo, W Shi, R Zhang, L Song, M Zhan, and H Mathcoder Li. Seamless code integration in llms for enhanced mathematical reasoning. *arXiv preprint arXiv:2310.03731*, 2023.

[33] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

[34] Elliot Nelson, Georgios Kollias, Payel Das, Subhajit Chaudhury, and Soham Dan. Needle in the haystack for memory based large language models. *arXiv preprint arXiv:2407.01437*, 2024.

[35] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024.

[36] Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. Kobbq: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 11:507–524, 2024.

[37] Jiho Jin, Woosung Kang, Junho Myung, and Alice Oh. Social bias benchmark for generation: A comparison of generation and qa-based evaluations. *arXiv preprint arXiv:2503.06987*, 2025.

[38] Jingnan Zheng, Xiangtian Ji, Yijun Lu, Chenhang Cui, Weixiang Zhao, Gelei Deng, Zhenkai Liang, An Zhang, and Tat-Seng Chua. Rsafe: Incentivizing proactive reasoning to build robust and adaptive llm safeguards. *arXiv preprint arXiv:2506.07736*, 2025.

[39] Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *ICLR 2024 (Spotlight)*, 2024.

[40] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[41] BIG-bench Collaboration. Bbh (big-bench hard): A subset of challenging tasks from big-bench. Dataset by BIG-bench project, 2022.

[42] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems: Gsm8k benchmark. In *NeurIPS 2021*, 2021.

[43] Mayank Aggarwal, Kirti Jain, Anil Kumar, Amanpreet Singh, Rahul Gupta, Parag Grover, V S Kanchana, Vipul Saini, Rishabh Jain, Anish Kumar, et al. Mbpp+: A diverse and challenging dataset for benchmarking code generation. *arXiv preprint arXiv:2303.04475*, 2023.

[44] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[45] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024.

[46] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

[47] LG AI Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Woohyung Lim, Sangha Park, Sooyoun Park, Sihoon Yang, Heuiyeen Yeen, and Hyeongu Yun. Exaone 3.5: Series of large language models for real-world use cases, 2024.

[48] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023.

[49] LG AI Research. Exaone 3.0 7.8b instruction-tuned language model, 2024. Technical Report.

[50] Won Ik Park, Jihwan Lee, Jinseok Seol, Beomsu Kim, Sangwoo Seo, Seongjin Park, Jihyung Moon, Sungdong Kim, Chan Young Park, Minjoon Seo, and Joongbo Shin. Kobest: Korean balanced evaluation of significant tasks, 2023.

[51] Based on Sakaguchi et al. (2021) and Open Ko-LLM Leaderboard Team. Ko-winogrande: Korean adaptation of winogrande commonsense reasoning benchmark. Included in Open Ko-LLM Leaderboard2, 2024. Referenced in arXiv:2410.12445, related to dataset adaptation.

[52] Kaijing Ma, Xinrun Du, Yunran Wang, Haoran Zhang, Zhoufutu Wen, Xingwei Qu, Jian Yang, Jiaheng Liu, Minghao Liu, Xiang Yue, Wenhao Huang, and Ge Zhang. Kor-bench: Benchmarking language models on knowledge-orthogonal reasoning tasks, 2024.

[53] Hyunwoo Ko, Guijin Son, and Dasol Choi. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap, 2025.

[54] Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. Kmmlu: Measuring massive multitask language understanding in korean. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) Long Papers*, pages 4076–4104, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[55] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.

[56] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers)*, pages 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[57] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

[58] Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking. *Advances in Neural Information Processing Systems*, 37:129696–129742, 2024.

8 Contributor

Within each role, names are listed in alphabetical order by first name. The first line is the leader of each group.

Pre-training

Hwijung Ryu
Changwon Ok
Hoyoun Jung
Hyesung Ji
Jeehyun Lim
Jehoon Lee
Ji-Eun Han
Jisoo Baik
Mihyeon Kim
Riwoo Chung
Seongmin Lee
Wonjae Park
Yoonseok Heo
Youngkyung Seo

Post-training

Seyoun Won
Boeun Kim
Cheolhun Heo
Eunkyeong Lee
Honghee Lee
Hyeongju Ju
Hyeontae Seo
Jeongyong Shim
Jisoo Lee
Junseok Koh
Junwoo Kim
Minho Lee
Minji Kang
Minju Kim
Sangha Nam
Seongheum Park
Taehyeong Kim

Engineering

Euijai Ahn
Hong Seok Jeung
Jisu Shin
Jiyeon Kim
Seonyeong Song
Seung Hyun Kong
Sukjin Hong
Taeyang Yun

Model Evaluation

Yu-Seon Kim
A-Hyun Lee
Chae-Jeong Lee
Hye-Won Yu
Ji-Hyun Ahn
Song-Yeon Kim
Sun-Woo, Jung

Data Sourcing

Eunju Kim
Eunji Ha
Jinwoo Baek
Yun-ji Lee

Responsible AI

Wanjin Park
Jeong Yeop Kim
Eun Mi Kim
Hyoung Jun Park
Jung Won Yoon
Min Sung Noh
Myung Gyo Oh
Wongyoung Lee
Yun Jin Park

Supportive role

Young S. Kwon
Hyun Keun Kim
Jieun Lee
YeoJoo Park

Director

Donghoon Shin (Model)
Sejung Lee (Data & Evaluation)
Soonmin Bae (RAI)