HW1: On the Opportunities and Risks of Foundation Models

Eunhwi Lee (lehlsy0904@snu.ac.kr)
Computational Clinical Science Laboratory, Seoul National University

Foundation models (FMs) mark a qualitative departure from prior machine and deep learning approaches. While earlier systems required domain-specific features or supervised objectives, FMs rely on large-scale self-supervision, enabling emergent abilities such as in-context learning, zero-shot transfer, and multimodal reasoning—capabilities that were not explicitly engineered but “emerged”. This shift illustrates a new paradigm where performance arises less from handcrafted design and more from scale and generalization.

Compared with earlier ML/DL models, FMs consolidate methods across domains with a single pretrained model, creating a homogenized paradigm that can be fine-tuned or prompted for almost any task. This broad adaptability raises both opportunities and risks: while improvements transfer widely, flaws and biases propagate just as easily.

Since the original release of this report, FMs have rapidly expanded in capability and accessibility. ChatGPT and GPT-4 brought FMs into everyday use, multimodal models now integrate text, vision, and speech, and debates on safety, alignment, and regulation have become central to AI governance.

For psychologists, FMs enable the integration of diverse data streams, from questionnaires and clinical interviews to physiological signals and ecological momentary assessment data. Such multimodal synthesis supports the development of AI-guided personalized interventions that reflect individualized phenotypes of mental disorders.