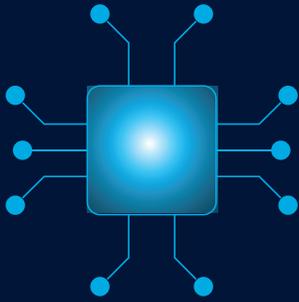


MODELS ARE PERISHABLE; DATA AND KNOWLEDGE ARE PERMANENT

There are many generative AI models available, and they are constantly evolving. A focus on data and tailoring open-source AI to specific domains and cultures is the best way to achieve business value at scale.



The dawn of OpenAI heralded a new epoch in the realm of artificial intelligence (AI). In January 2024, firms had access to **over 475,000 closed- and open-source foundation models**. This is an indication of the amount of data and possibilities, and why it is not easy to decide which model to choose.

The Problem with Models

Selecting the appropriate model – whether closed or open – requires tailoring the choice to different enterprise scenarios and use cases. Enterprises need to know that the models chosen are commercially robust, offer customization of features and agent capabilities, and provide rigorous enterprise-grade security and privacy. They also need to choose models that align with responsible AI tenets, keep up with the changing regulatory landscape, and offer tamper-resistant watermarks to build trust with consumers.

While closed foundational models are at the forefront of enterprise adoption, they can be costly and, in many instances, inflexible. The most effective method to achieve domain-centricity is through retrieval augmented generation (RAG). However, these models can be challenging to fine-tune or customize, and are often prohibitively expensive. Open-source models hold the key to calibration, empowering enterprises to adapt them to specific requirements, resulting in superior outcomes and heightened efficiency. Nevertheless, adoption remains weak due to commercial use restrictions and an absence of necessary features.

Thinking Ahead

With robust hardware and training methods, however, we will witness the emergence of open foundation models that are both permissible and affordable.

In light of this shifting landscape, the following considerations are key:

1. Data over models

As the number of models proliferates over time, they also become outdated or irrelevant. Illustratively, a leading energy corporation transitioned from GPT-3.x, citing inadequacy in memory and reasoning capabilities when compared with GPT-4.0. Data and knowledge in any organization have enduring value. Hence, organizing data in a model-agnostic manner becomes essential to realize business benefits. This strategy facilitates easy access and management of data, regardless of the model employed. Advantages include the creation of centralized data repositories that can be utilized by various models, and data sharing across models to foster improved collaboration and efficiency across the enterprise.

2. Affordable AI

Developing and training AI models comes at an expense that could make even the most eager leader lose heart. However, the growing availability of models means that most firms can find models that execute a wide range of tasks with high precision and at low expense. This transition towards cost-effective AI will allow small and medium enterprises (SMEs) to embrace AI solutions, resulting in enhanced productivity and efficiency.

3. Fine-tuning for domain-specificity

Adjusting models to work within specific domains is vital to ensure industry-specific tasks and use cases can be onboarded effectively. This method has proven successful at Infosys, where models fine-tuned for specific domains have resulted in increased accuracy and efficiency. For example, in one case of fine-tuning using open-source code large language model (LLM), the task took us four weeks for five million lines of application code. The fine-tuned model was 20% more accurate than the generic GitHub Copilot. Evidently, fine-tuning allows AI to address specific needs and challenges, yielding superior outcomes and heightened customer satisfaction.

4. Importance of local data and diverse languages

AI models trained on local data and diverse languages onboard the subtleties of local cultures and linguistic nuances, resulting in increased model accuracy and efficacy. Nandan Nilekani, Infosys Chairman, advocates the importance of tailoring AI models to specific regions and cultures: "The data is what is missing: we don't have Indian language data and we don't have local data. So, fix the data problem," Nilekani said while referring to India's DPI roadmap.

Expectations are High

The future of AI appears bright, with open-source models becoming both permissible and accessible to all. This shift towards open source will foster increased collaboration and innovation, as developers both build upon and enhance existing models.

Humans will play a central role in fortifying the relationship between data and generative AI, serving as the linchpin in this ongoing cycle of reinforcement. The focus on data over models will continue to expand. Firms that organize their data in a model-agnostic manner, develop affordable AI models, fine-tune models for specific domains, and consider local data and diverse languages in the model training phase are the ones that will take generative AI from hype to value-driver in the years ahead.



We have curated the [top 10 AI imperatives](#) from our own learnings and experience into Infosys Topaz, our AI-first set of services, solutions and platforms using generative AI technologies. With 12,000+ AI assets, 150+ pre-trained AI models, 10+ AI platforms steered by AI-first specialists and data strategists, and a 'responsible by design' approach, Infosys Topaz helps enterprises accelerate growth, unlock efficiencies at scale and build connected ecosystems. Connect with us at infosystopaz@infosys.com.

For more information, contact askus@infosys.com



© 2024 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.