Huawei's ambitions for GenAI go beyond China, and beyond telecoms

October 2023

Martin Scott

Huawei, as the world's largest telecoms equipment vendor by revenue, is diversifying, focusing not only on protecting and developing its existing telco business lines but also on deepening its relationship directly with enterprises. The Huawei Connect 2023 event was the vendor's opportunity to showcase its vision for AI to both audiences.

Huawei is working to develop an all-Chinese solution for generative AI (GenAI)

Several factors incentivised Huawei to pursue a GenAI strategy focused on an all-Chinese value chain.

Firstly, the USA has blocked the export of chip-making technology to China since 2019. This restricted the domestic manufacturing of the higher-powered GPUs (Graphical Processing Units) and NPUs (Neural Processing Units) necessary to train large language models (LLMs) efficiently, but Chinese industrial processes have evolved.

For example, Huawei's most-recently-released flagship handset, the Mate 50 Pro, is powered by a 7nm-engraved processor. Even though 3nm processors are now becoming mainstream (as used in Apple's A17 Pro processor), the 7nm engraving process was previously considered to be impossible without access to USA proprietary technology.

Secondly, China's economy is stagnating – manufacturing activity is contracting and the property bubble has burst. GenAI-driven efficiency is important for future growth. All Chinese tech giants are developing foundational models and Huawei is a key domestic supplier for the enabling hardware.

China is developing its own silicon, its own foundational models (Huawei's is called Pangu – it is optimised for both English and Chinese languages and runs on Chinese processors) and will be supplying solutions to its domestic market. Huawei Connect 2023 was primarily aimed at government and domestic enterprise buyers, convincing them to buy into Huawei's vision of AI and the potential efficiencies that Huawei's solutions might unlock. Huawei Cloud, the division of Huawei that is responsible for much of the AI initiative, is therefore positioning itself as an alternative to GenAI solutions from companies like AWS and Microsoft.

Huawei's plans to offer inference-as-a-service are credible

GenAI foundational models are expensive to train, but a large number of Chinese businesses are doing so anyway. According to brokerage firm CLSA over 130 Chinese LLMs had been created to September 2023 – 40% of the worldwide total. The domestic demand is unlikely to support so many models and only a select handful will produce commercially viable businesses. Huawei's Pangu is likely to be one such model that remains viable due to both Huawei's scale and due to Huawei's focus on efficient deployment.

Foundational models are typically general-purpose and need either a second round of training ('fine-tuning') to add further knowledge or the support of a range of ancillary tools that interpret prompts and add context before they are passed to the LLM.

There are two main groups of such tools. The first are called 'agents'. Agents adjust the prompts and add context so that the foundational model produces more relevant outputs. The second are the Retrieval Augmented Generation (RAG) frameworks. These frameworks allow the AI access to data lakes of relevant data for their specific use case. The majority of present GenAI solutions from AWS, Microsoft and others, typically take a two-step approach to training AI: first they train (or take) a foundational model, fine-tune it, and then rely on agents and RAG to tailor the AI to specific use cases. Huawei has generalised a three-step model:

- train foundational model
- fine-tune with sector knowledge (for example, mining, telecoms)
- fine-tune with application knowledge (for example, call centre support, network optimisation).

After this, similar agent and RAG assistance is also applied.

This three-tier training method has two benefits. Firstly, Huawei can very rapidly deploy off-the-shelf inference-as-a-service solutions for applications and industries for which it has already tailored its Pangu model.

Secondly, this three-step process potentially leads to more energy-efficient models than those that are trained in a two-step process because unnecessary bits of the model are pruned during each fine-tune. That leads to a lower dependency on agents to interpret prompts which makes the model cheaper to run. Huawei and its partners need their solutions to be more energy- (and cost-) efficient than GPT-based alternatives because, on many metrics, they do not yet match the performance of GPT-4.

Overall, though, Huawei's messaging around AI, and its approach – to move sector by sector from less complex and more-predictable environments (for example, automotive manufacturing) up to higher complexity (mining, healthcare) was well-considered and well-articulated.

Huawei is looking beyond telecoms for GenAI opportunities, but telecoms remains a core competency area

Huawei is actively looking beyond the telecoms sector for future growth. It sees GenAI as a vehicle not just to promote its hardware, software and cloud solutions but also to promote the connectivity that necessarily underpin these services to a wider enterprise audience. Industrial efficiency is at the core of the message, albeit with limited acknowledgement of the importance of retaining the workforce (youth unemployment in China is reportedly at 21%, the highest level in recent history).

GenAI will be applied transformatively to replace human labour in parts of the Chinese economy and in other regions where Huawei retains its well-established presence, such as in Latin America and emerging countries in Asia. Regulation is less likely to slow deployment in these regions in the near future compared to Europe and North America. Here, Huawei is directly positioning Huawei Cloud against the likes of AWS, Microsoft and Google as a 'second option for the world'. Many of these markets may buy Huawei solutions for GenAI applications but telecoms, in particular, is the sector where Huawei has greater experience than the hyperscalers and is where it is arguably still most likely to succeed in gaining traction for its GenAI propositions.