Fabric 1.6

System Card

Fabric AI

research@fabricai.co.uk

Version 1.0 | August 3, 2026

Contents

1 Introduction

Fabric 1.6 is a 35-billion-parameter Mixture-of-Experts (MoE) reasoning model developed by Fabric AI, with approximately 3 billion parameters active for each generated token. It combines a hybrid linear-attention architecture with explicit reasoning: Fabric 1.6 produces an internal chain of thought before answering, allowing it to plan, verify and correct its own reasoning on difficult tasks.

The model is designed for *agentic* use in harnesses such as OpenCode, Pi Agent, Hermes Agent and other OpenAI-compatible tool-calling environments. To support long-running, multi-turn agentic sessions, Fabric 1.6 offers the option to preserve thinking context from past messages, so that reasoning from earlier turns remains available to later ones.

Key characteristics of Fabric 1.6:

- **Native long context.** A context window of 262,144 tokens natively, extensible to 1,010,000 tokens.
- **Multi-Token Prediction (MTP).** A dedicated MTP module predicts multiple future tokens per step, enabling up to 50% faster generation during inference.
- **Multimodal input.** Fabric 1.6 accepts both text and image inputs, enabling vision-grounded reasoning and agentic tasks.
- **Reasoning transparency.** The model exposes its reasoning in a structured format that can be streamed, stored and reasoned over by the surrounding system.

This system card describes the model architecture (Section 2), the data and training programme (Section 3), evaluation results (Section 4), intended use (Section 5) and known limitations (Section 6).

2 Model Architecture

Fabric 1.6 is a hybrid attention model in which gated linear-attention (DeltaNet) layers and gated full-attention layers alternate within a Mixture-of-Experts (MoE) transformer. Table 1 summarises the architecture.

Component	Specification
Parameters	35B total; ≈3B active per token
Hidden dimension	2,048
Layers	40
Layer layout	$10 \times (\, 3 \times$ (Gated DeltaNet $\rightarrow$ MoE) $\rightarrow 1 \times$ (Gated Attention $\rightarrow$ MoE) $)$
Gated DeltaNet	32 linear value heads and 16 linear QK heads, head dimension 128
Gated Attention	16 query heads and 2 key/value heads, head dimension 256; rotary embedding dimension 64
Mixture of Experts	256 experts; 8 routed + 1 shared expert per token; expert intermediate dimension 512
Context length	262,144 tokens native; extensible to 1,010,000
Multi-Token Prediction	1 MTP layer; up to 50% faster generation

Table 1: Architecture summary of Fabric 1.6.

The gated DeltaNet layers provide efficient, hardware-friendly linear attention that scales sub-quadratically with sequence length, while the regularly interspersed gated attention layers

(one per group of four) provide full attention capacity where it matters most. This hybrid design is what enables the model's 262,144-token native context length without quadratic-cost attention at every layer.

3 Model Data and Training

3.1 Base model

Fabric 1.6 is built on *Qwen3.5-35B-A3B-Base* [1], an open 35-billion-parameter MoE base model with approximately 3 billion active parameters, released by the Qwen team (Alibaba). From this foundation, Fabric 1.6 was developed through a two-stage training programme: *continuous pre-training*, followed by *post-training* comprising supervised fine-tuning (SFT) and reinforcement-learning-based alignment. All stages were conducted with rigorous data filtering, deduplication and quality control on an **NVIDIA DGX A100 8-GPU Server System with 320GB VRAM**.

3.2 Continuous Pre-Training

Fabric 1.6 was continuously pre-trained primarily on a large, in-house proprietary synthetic dataset spanning code, mathematics and reasoning, complemented by open reasoning corpora:

- **OpenThoughts3-1.2M** [2] — 1.2M high-quality reasoning traces spanning mathematics, science, coding and general problem solving.
- **OpenR1-Math-220k** [3] — 225k mathematical problems with think-style solutions.

In total, approximately **12 billion tokens** were processed across the pre-training and post-training stages. The proprietary synthetic corpus is generated in-house by Fabric AI and is designed to reinforce structured reasoning, domain knowledge and instruction-following behaviour before instruction tuning begins. Knowledge cutoff: July 2026.

3.3 Post-Training: Supervised Fine-Tuning

The continuously pre-trained model was instruction-tuned on a diverse, high-quality SFT corpus combining open, permissively licensed data with proprietary data:

- **smoltalk2** (Apache-2.0 SFT subset) [4] — an approximately 340k-example subset covering multilingual instruction following, multi-turn reasoning, tool-calling traces, system chats and table understanding.
- **hermes-function-calling-v1** [5] — structured tool-calling traces for reliable function invocation.
- **oasst2** [6] — curated, reviewed conversational chains.
- **maple** — a proprietary instruction and reasoning corpus developed in-house by Fabric AI.

3.4 Post-Training: Reinforcement Learning and Alignment

Following supervised fine-tuning, the model was aligned through reinforcement learning from feedback, optimising for the behaviours that matter most in agentic deployment: instruction following, reliable tool-calling, appropriate reasoning depth, helpfulness and calibrated refusals.

This stage is designed to improve the model's behaviour end-to-end — including how it decides *whether* to reason, how long it reasons, and how it acts on the result — rather than optimising a single proxy objective.

3.5 Data Governance

All training data is subject to automated filtering and deduplication. Personal information is removed where identifiable, and permissively licensed (Apache-2.0) open datasets are retained and attributed in accordance with their licences (see References). Proprietary corpora are governed under Fabric AI's internal data policies.

4 Evaluation

Fabric 1.6 was evaluated on 22 benchmarks spanning mathematics, science and knowledge, coding, general reasoning and agentic tool use. Table 2 reports the results.

Methodology. All evaluations were run with greedy decoding (temperature 0) and a fixed random seed. Self-contained suites (AIME, HMMT, IMO AnswerBench, MATH-500, GSM8K Platinum) were evaluated with the native model implementation; code and agentic suites (LiveCodeBench v6, SWE-bench Verified, SWE-bench Pro, TAU3-Bench, MCP-Atlas, WideSearch) were run through an OpenAI-compatible serving endpoint with their official harness packages; multimodal suites (MMMU-Pro, RealWorldQA, MathVista mini) used image inputs.

Benchmark	Metric	Samples	Score
Math & Reasoning			
AIME 2025	pass@1	30	92.8
AIME 2026	pass@1	30	93.1
HMMT February 2026	pass@1	44	83.2
IMO AnswerBench	pass@1	520	79.2
MATH-500	accuracy	500	84.8
Science & Knowledge			
GPQA	accuracy	198	86.7
GPQA Diamond	accuracy	198	84.9
Humanity's Last Exam	accuracy	2,500	21.4
MMLU-Pro	accuracy	12,032	85.6
MMLU-Redux	accuracy	3,000	93.5
C-Eval	accuracy	1,346	92.3
Coding			
LiveCodeBench v6	pass@1	500	80.2
SWE-bench Verified	resolve rate	500	72.9
SWE-bench Pro	resolve rate	1,200	50.1
IFEval	instruction-level	541	93.09
General Reasoning			
GSM8K Platinum	accuracy	1,319	95.73
Agentic Tools			
TAU3-Bench	pass rate	128	67.2
MMMU-Pro	accuracy	1,260	74.10
RealWorldQA	accuracy	765	85.4
MCP-Atlas	completion	64	62.8
WideSearch	rubric score	100	60.3
MathVista mini	accuracy	1,000	86.6
Mean (all 22 benchmarks)			**78.5**

Table 2: Fabric 1.6 evaluation results across 22 benchmarks.

5 Intended Use

Fabric 1.6 is intended for agentic and reasoning-heavy workloads:

- **Agent harnesses.** OpenCode, Pi Agent, Hermes Agent and other tool-calling environments, including OpenAI-compatible API serving.
- **Long-context reasoning.** Multi-turn analysis of large documents, codebases and conversation histories, with optional preservation of thinking context across messages.
- **On-premises deployment.** The model is distributed in BF16 and GGUF formats and runs on CPU/GPU inference stacks, including llama.cpp-based servers with flash attention, quantised KV caches and MTP-assisted speculative decoding.

Fabric 1.6 is *not* intended for use without output-level safeguards in high-stakes domains (see Section 6).

6 Safety and Limitations

As with all large language models, Fabric 1.6 can produce plausible but incorrect information, particularly in niche or rapidly changing domains. Reasoning improves accuracy but does not guarantee factual correctness; verification against authoritative sources is recommended for high-stakes use.

In agentic settings, the model's outputs should be treated as untrusted and executed only within sandboxed, permissioned environments. Tool-calling behaviour, while optimised through post-training, remains probabilistic and should be paired with schema-level validation and human oversight where actions have real-world consequences.

Fabric 1.6 was primarily evaluated on the benchmarks in Section 4. Its behaviour across all languages, cultures and demographic groups has not been exhaustively measured; bias and robustness testing is ongoing. The model may exhibit limitations in very long contexts beyond its native window, and reasoning depth varies with task difficulty.

7 Contact

For questions, collaborations or access requests, contact the Fabric AI research team at research@fabricai.co.uk.

References

[1] Qwen Team (Alibaba). *Qwen3.5-35B-A3B-Base*. https://huggingface.co/Qwen/Qwen3.5-35B-A3B-Base

[2] Open Thoughts. *OpenThoughts3-1.2M*. https://huggingface.co/datasets/open-thoughts/OpenThoughts3-1.2M

[3] Open R1. *OpenR1-Math-220k*. https://huggingface.co/datasets/open-r1/OpenR1-Math-220k

[4] HuggingFaceTB. *smoltalk2*. https://huggingface.co/datasets/HuggingFaceTB/smoltalk2

[5] NousResearch. *hermes-function-calling-v1*. https://huggingface.co/datasets/NousResearch/hermes-function-calling-v1

[6] OpenAssistant. *oasst2*. https://huggingface.co/datasets/OpenAssistant/oasst2