MODEL CARD June 2026 v2.1.0

# Lumina-7B-Instruct

TensorForge AI • Causal Language Model • 7.3B Parameters • 32K Context

## Model Details

**Summary.** Lumina-7B-Instruct is a 7.3-billion-parameter decoder-only Transformer language model, fine-tuned via supervised instruction tuning and RLHF for general-purpose conversational and coding tasks. It supports a context window of 32,768 tokens and is released under the TensorForge Community License.

| | |
|---|---|
| **Organization** | TensorForge AI (tensorforge.ai) |
| **Model Type** | Causal Language Model — decoder-only Transformer |
| **Version** | 2.1.0 (3 July 2026) |
| **License** | TensorForge Community License v1.0 |
| **Parameters** | $7.3 \times 10^9$ (7.3B) |
| **Architecture** | 32 layers, 4096 hidden size, 32 attention heads, Grouped-Query Attention (4 KV heads), RoPE embeddings, SwiGLU activations |
| **Tokenizer** | SentencePiece BPE, vocabulary size 128 000 |
| **Context Length** | 32 768 tokens (input + output) |
| **Training Data** | Mix of filtered web text (CommonCrawl), books, Wikipedia, GitHub code, and curated instruction datasets. Knowledge cutoff: December 2025. |
| **Input / Output** | Text in, text out. Supports system prompts and multi-turn conversation. |
| **Languages** | English (primary); 30+ languages at varying proficiency |
| **Hardware** | 256 × Lumina H100-80GB GPUs; 18 days training |
| **Precision** | BF16 training; FP16 inference |
| **Fine-Tuning** | Supervised fine-tuning + RLHF (PPO) on 180K human preference pairs |
| **Contact** | safety@tensorforge.ai • tensorforge.ai/models/lumina |

## Intended Use

Lumina-7B-Instruct is designed as a general-purpose assistant for **non-high-stakes** applications. Below is the intended-use matrix evaluated against our internal safety and quality rubrics.

| Use Case | Suitability | Notes |
|---|---|---|
| General Q&A | High | Core strength. Factual accuracy varies by domain. |
| Code Generation | High | Strong on Python, JavaScript, TypeScript, Go. Weaker on niche DSLs. |
| Creative Writing | Medium | Good fluency; occasional repetitiveness in long-form. |
| Summarisation | High | Reliable for documents up to 20K tokens. |
| Mathematics | Limited | Struggles with multi-step reasoning. Verify outputs. |
| Multilingual | Medium | Best for high-resource languages (Spanish, French, German, Japanese). |
| Medical Advice | Not Recommended | Not validated for clinical use. High hallucination risk. |
| Legal Analysis | Not Recommended | Not trained on jurisdiction-specific case law. |
| Content Moderation | Limited | Requires external guardrails for production filtering. |
| Customer Support | Medium | Suitable with human-in-the-loop oversight. |

See Section 6 for detailed limitations and safety guidance.

## Evaluation Results

We evaluate Lumina-7B-Instruct on standard academic benchmarks. Scores are colour-coded by performance tier: $\geq 80$ 70–79 60–69 50–59 $< 50$.

| Benchmark | Score | Tier | Category |
|---|---|---|---|
| MMLU (5-shot) | 74.2 | 70–79 | World Knowledge |
| HellaSwag (10-shot) | 84.6 | $\geq$80 | Commonsense Reasoning |
| ARC-Challenge (25-shot) | 72.1 | 70–79 | Science Reasoning |
| WinoGrande (5-shot) | 78.9 | 70–79 | Coreference Resolution |
| TruthfulQA (0-shot) | 65.3 | 60–69 | Factuality & Safety |
| GSM8K (8-shot, CoT) | 58.7 | 50–59 | Mathematical Reasoning |
| MATH (4-shot) | 38.4 | <50 | Advanced Mathematics |
| HumanEval (0-shot) | 73.8 | 70–79 | Code Generation (Python) |
| MBPP (3-shot) | 68.1 | 60–69 | Code Generation (Python) |
| MultiPL-E (JS, 0-shot) | 71.5 | 70–79 | Multilingual Code |
| BBH (3-shot, CoT) | 66.2 | 60–69 | Challenging Reasoning |
| AGIEval (English, 0-shot) | 63.9 | 60–69 | Standardised Exams |

**Note.** Few-shot settings follow EleutherAI LM Harness conventions. Chain-of-thought (CoT) prompting is denoted where used. Scores are the mean over 3 random seeds ($\sigma < 0.8$ across all benchmarks).

## Fairness Analysis

We assess performance parity across demographic axes using the **BBQ** (Bias Benchmark for QA) and **WinoBias** datasets. The charts below report accuracy disaggregated by subgroup.

| BBQ Subgroup | Acc. | BBQ Subgroup | Acc. |
|---|---|---|---|
| Gender – Female | 74.1 | Gender – Male | 78.3 |
| Gender – Non-binary | 71.6 | Race – White | 80.2 |
| Race – Black | 69.8 | Race – Asian | 77.5 |
| Race – Hispanic | 72.3 | Age – 18–30 | 76.4 |
| Age – 31–50 | 79.1 | Age – 51+ | 73.9 |
| Religion – Christian | 78.6 | Religion – Muslim | 67.2 |
| SES – High | 81.0 | SES – Low | 70.5 |

**Key observations.** Performance gaps are most pronounced along the Religion ($\Delta = 11.4$ pp, Muslim vs. Christian) and Socio-economic ($\Delta = 10.5$ pp, Low vs. High) axes. We recommend additional de-biasing fine-tuning and targeted evaluation before deploying in contexts where these disparities could cause harm.

## Example Usage

Listing 1: Inference with the Transformers library.

```
from transformers import AutoModelForCausalLM, AutoTokenizer

model_id = "tensorforge/Lumina-7B-Instruct"
tokenizer = AutoTokenizer.from_pretrained(model_id)
model = AutoModelForCausalLM.from_pretrained(
    model_id,
    torch_dtype="auto",
    device_map="auto",
)

messages = [
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "Explain backpropagation in 3 sentences."},
]
inputs = tokenizer.apply_chat_template(
    messages, tokenize=True, return_tensors="pt"
).to(model.device)

outputs = model.generate(inputs, max_new_tokens=256, temperature=0.7)
print(tokenizer.decode(outputs[0], skip_special_tokens=True))
```

## Caveats & Limitations

### ⚠ Critical Caveats

- **Hallucination.** Like all LLMs, Lumina-7B-Instruct may generate plausible-sounding but factually incorrect content. Always verify outputs in high-stakes settings.
- **Bias & Stereotypes.** Training data reflects web-scale biases. The model may reproduce stereotypes, particularly for under-represented groups (see Fairness Analysis, Section 4).
- **Out-of-Scope Use.** This model has *not* been validated for medical diagnosis, legal advice, safety-critical systems, or decisions with material impact on individuals' rights.
- **Adversarial Robustness.** The model is vulnerable to prompt injection, jailbreaking, and extraction attacks. Deploy with input/output filtering in production.
- **Privacy.** Training data was filtered for PII, but residual memorisation of rare sequences is possible. Do not expose the model to sensitive user data unless additional privacy-preserving measures are in place.
- **Multilingual Fairness.** Performance degrades significantly for low-resource languages. Users should evaluate independently for their target languages.

## Recommendations for Downstream Use

- Perform **domain-specific evaluation** on your target task and demographic distribution before deployment.
- Apply **output guardrails** (toxicity classifiers, factuality checkers) in user-facing applications.
- Monitor **drift and feedback loops** in production—model behaviour can shift with changing usage patterns.
- Where feasible, provide users with **transparency mechanisms**: confidence scores, source attribution, or an opt-out for AI-generated content.

## How to Cite

Listing 2: BibTeX citation.

```
@misc{tensorforge2026lumina,
  title    = {Lumina-7B-Instruct: A 7.3B Instruction-Tuned Language Model},
  author   = {TensorForge AI},
  year     = {2026},
  month    = jul,
  version  = {2.1.0},
  url      = {https://tensorforge.ai/models/lumina},
  note     = {Model Card},
}
```

tensorforge.ai/models/lumina  github.com/tensorforge/lumina  safety@tensorforge.ai