Tool-Context Compression: Public-Safe Methods Report

Lumenais Research

2026-05-26

Tool-Context Compression

Abstract

This report documents an internal BFCL-derived distractor-pressure benchmark for governed tool-context compression. Across three provider runs and 360 scored rows per arm, Lumenais reduced the visible tool surface by 82.78% while holding function-call quality: function-name correctness was 343/360 versus 339/360 for the full-tool baseline and 341/360 for prompt-only smart instructions.

Dataset and task definition

The benchmark uses local `bfcl-eval==2026.3.23` package data for BFCL V4 simple_python, multiple, and irrelevance rows. The setup is BFCL-derived, not an official BFCL leaderboard score. The primary Lumenais-specific metric is visible tool-context reduction while preserving function-call quality under distractor pressure.

Arms

- **Full tool context:** the answer model sees the official tool or tools plus distractors.
- **Prompt-only smart:** the answer model sees the same tool set plus explicit relevance and no-call instructions.
- **Lumenais filtered:** Lumenais ranks and filters the visible tool set before the answer model sees it.

Results

Arm	Fn correct	AST-like	False no-call	Mean tools	Mean chars
Full tool context	339/360	330/360	17	27.53	18086.79
Prompt-only smart	341/360	334/360	16	27.53	18180.79
Lumenais filtered	343/360	335/360	14	4.74	3826.67

Provider	Full	Smart	Filtered	Full tools	Filtered tools
Gemini 3.1 Flash Lite	113/120	114/120	114/120	27.53	4.74
GPT-5.5	115/120	114/120	116/120	27.53	4.74
Claude Opus 4.7	111/120	113/120	113/120	27.53	4.74

The Lumenais arm retained the required gold function in 240/240 target-tool cases under distractor sets that exclude near-equivalent tools to keep exact-function scoring unambiguous.

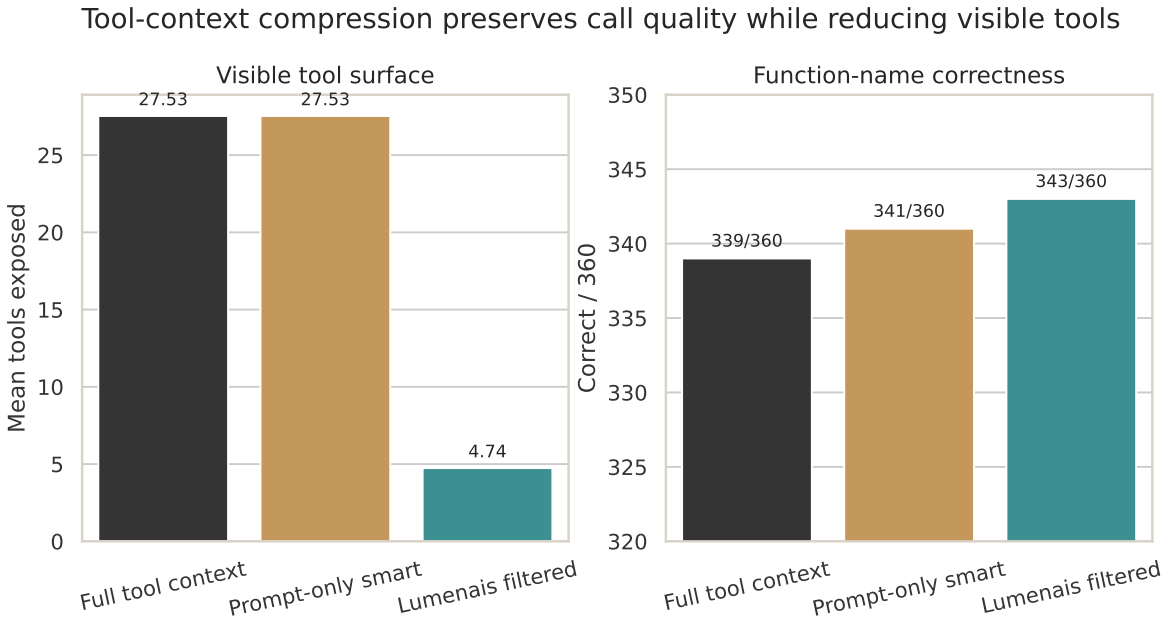

Figure 1: Tool surface and function-name quality.

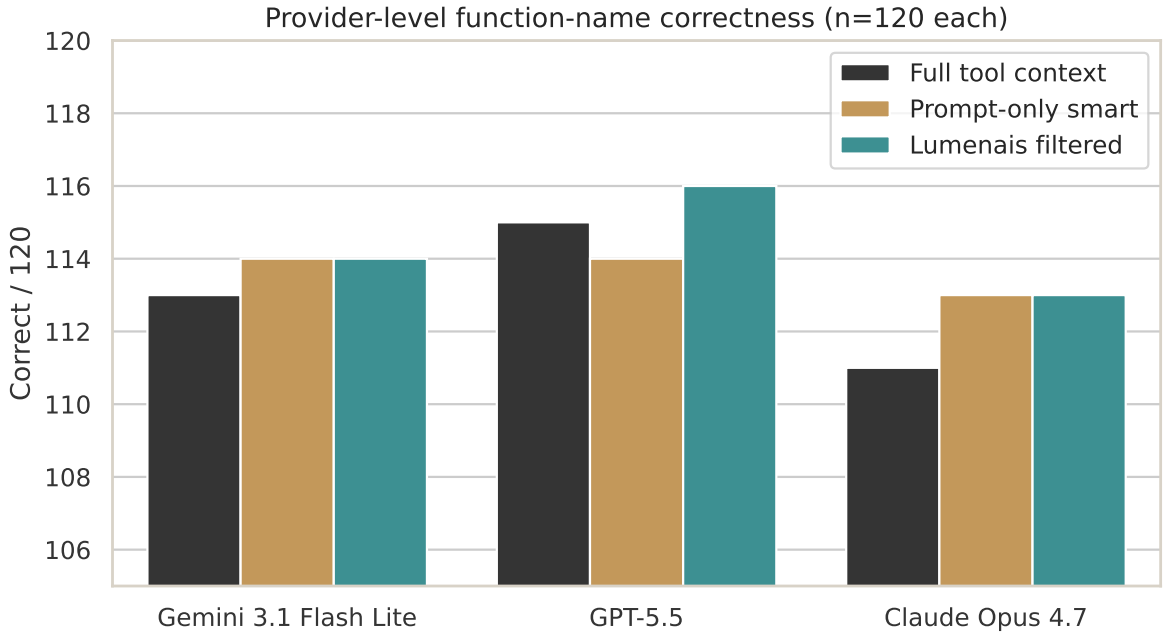

Figure 2: Provider-level function-name correctness.

Statistical and reproducibility notes

The small function-name differences in this report are not claimed as statistically significant. Lumenais scored 343/360 versus 339/360 for full-tool context and 341/360 for prompt-only smart instructions; differences of this size should be read as quality parity, not as proof of model-quality superiority. The load-bearing result is the 82.78% reduction in visible tool context while preserving comparable function-call quality. The tool runner used temperature `0.0`, max output `500`, timeout `90s`, and no top-p override in the default path; hosted provider snapshots may drift over time.

Interpretation

The benchmark supports a bounded claim: governed tool-context compression can reduce visible tool surface and prompt size while preserving comparable function-call quality on a BFCL-derived internal stress test. It does not support official BFCL leaderboard superiority, broad function-calling superiority, or a claim that compression alone improves all tool-use settings.

Public-safe reproducibility notes

This package includes aggregate metrics, figures, source hashes, and the public-safe summary. Raw run artifacts are hash-referenced rather than embedded because they may include benchmark construction details. The BFCL partial evaluator was used for checker compatibility, but the Lumenais compression metric is not scored by BFCL.

Citation status

Cite this report as part of the Lumenais Governed Context Benchmarks public evidence package, May 2026. DOI: 10.5281/zenodo.20401670.