HARVARDNETWORK / INDEPENDENT RESEARCH

GPT-5.5 Model Service Evaluation White Paper

A cross-validated assessment of capability, cost, hallucination, safety and deployment tradeoffs across GPT-5.5, Claude, Gemini and DeepSeek.

JULY 12, 2026

13 CHAPTERS / 45 EVIDENCE TABLES

GPT-5.5 Model Service Evaluation White Paper - Executive Summary

***Evaluation subject**: OpenAI GPT-5.5 (model service layer, API + ChatGPT interface)*
***Benchmark competitors**: Claude Opus 4.7 (Anthropic), Gemini 3.1 Pro / 3.5 Flash (Google), DeepSeek-V4 (DeepSeek)*
***Draft date**: 2026-07-12*

Independence Declaration

This report is an independent evaluation research synthesis with no commercial affiliation with OpenAI, Anthropic, Google, DeepSeek, or any other AI model vendor evaluated herein. No AI company provided financial support, computing resources, or pre-release access for this research. The author holds no equity in and has no consulting relationship with any company whose product is evaluated.

Evaluation evidence is drawn from three categories of source: independent third-party measurements (primarily Artificial Analysis), practitioner testing reports (independent benchmark runs and hands-on evaluations), and vendor-published benchmark claims (always labeled as vendor-reported and cross-checked against independent measurements wherever possible).

Funding source: [Institution Name to be inserted before publication]. Conflict of interest declaration: see Appendix A. All cited sources are publicly accessible; URLs and references are provided in each chapter's reference list.

Methodology Positioning

This white paper is a **meta-analysis and cross-validation** of publicly available evaluation evidence as of July 2026. The author did not execute a new 5,000-prompt API test campaign. Instead, the report synthesizes:

- **Independent measurements** (Artificial Analysis): AA Intelligence Index, AA-Omniscience, GDPval-AA
- **Practitioner testing reports**: documented benchmark runs and hands-on evaluations by independent practitioners (Berman, Ben Davis, Wes Roth, and additional industry technical reviewers)
- **Vendor-reported benchmark claims**: OpenAI's published benchmark scores and system card assertions, *always labeled as vendor-reported and cross-checked against independent measurements*

The original research plan called for a 5,000-prompt primary API test campaign; this report instead applies a meta-analysis methodology that synthesizes existing independent measurements and explicitly identifies which findings rest solely on vendor-reported figures. Chapter 11 (Limitations) quantifies the resulting evidence gaps.

Key Findings

Finding 1 - Intelligence Index ranking, independently verified

In the Artificial Analysis Intelligence Index - an independent third-party benchmark aggregating multiple cognitive tasks - GPT-5.5 scored **3 points higher** than the previous three-way tie among Anthropic, Google, and OpenAI flagships (Artificial Analysis, 2026).

On GDPval-AA, a 44-profession real-work evaluation hosted by Artificial Analysis, GPT-5.5 (xhigh) achieved **1,785 Elo**, approximately 30 points ahead of Claude Opus 4.7 (Max) and 470 points ahead of Gemini 3.1 Pro Preview (Artificial Analysis, 2026).

Finding 2 - Cost-efficiency lead at the Medium reasoning tier

At the **Medium** reasoning-effort setting, GPT-5.5 matched the intelligence of Claude Opus 4.7 (Max) at approximately **one quarter the cost** (~$1,200 vs ~$4,800 to run the AA Intelligence Index). GPT-5.5 (Low) approximated Claude Opus 4.7 (Non-reasoning, High) at approximately **one half the cost** (~$500 vs ~$1,000) (Artificial Analysis, 2026).

Across the full AA Intelligence Index, GPT-5.5 used **~40% fewer output tokens** than GPT-5.4 to reach the same score, absorbing most of the 2× per-token price increase. The net cost of running the Index rose by only **~20%**, and GPT-5.5 remained approximately **30% cheaper** than Claude Opus 4.7 (Max) (Artificial Analysis, 2026).

Finding 3 - Most severe weakness: hallucination rate

In the AA-Omniscience benchmark - which jointly measures knowledge recall and hallucination tendency - GPT-5.5 (xhigh) exhibited a **hallucination rate of 86%**, compared with **36% for Claude Opus 4.7 (Max)** and **50% for Gemini 3.1 Pro Preview** (Artificial Analysis, 2026). GPT-5.5 simultaneously achieved the **highest knowledge accuracy at 57%** across all tested models.

The interpretation: GPT-5.5 is more likely to produce a confident-sounding answer when uncertain, rather than declining to answer. This is the most consequential failure mode identified by independent measurement in this report.

Finding 4 - Vendor-reported benchmarks not fully reproduced

OpenAI reported a cluster of headline benchmark scores at the GPT-5.5 launch (OpenAI, 2026):

Benchmark	Vendor-reported score
Terminal-Bench 2.0	82.7%
SWE-Bench Pro	58.6%
Expert-SWE	73.1%
GDPval	84.9%
OSWorld-Verified	78.7%
CyberGym	81.8%
BrowseComp	84.4%
Tau2 Telecom	98.0%
FrontierMath Tier 1-3	51.7%

These figures have **not been independently reproduced at the same magnitudes** by independent measurement institutions as of the report date. Artificial Analysis confirmed GPT-5.5's leading position on Terminal-Bench Hard, GDPval-AA, and APEX-Agents-AA, but Gemini 3.1 Pro Preview led GPT-5.5 in **three additional evaluations** that AA hosts (Artificial Analysis, 2026). All vendor-reported figures in this report are flagged as "OpenAI-reported, pending independent verification."

Finding 5 - Specific failure modes documented by user testing

Independent user testing (Berman, 2026; an independent technical review, 2026) documented the following failure cases with full prompt-output evidence:

- **Multimodal OCR failure**: GPT-5.5 failed to recognize a regional university logo where Gemini 3.5 Flash succeeded (independent technical review, 2026).

- **Instruction-following failure on creative writing**: When asked for a 300-word sci-fi short story, GPT-5.5 produced 450-500 words, wrote a spy thriller instead of sci-fi, and handled the constraint "the male protagonist cannot speak" by literally hanging a "no speaking allowed" sign on him (independent technical review, 2026).
- **Long-context degradation**: Independent practitioner Ben Davis reported that GPT-5.5's context compaction is weaker than GPT-5.4, recommending thread length be kept under 100K tokens even though the API advertises a 400K-1M token window (Ben Davis, 2026).
- **Safety filter regression in GPT-5.5 Instant**: OpenAI's own system card reports gore-content filtering regressed from 0.867 to 0.703 and sexual-content filtering from 0.857 to 0.806 (lower is better); jailbreak resistance also regressed (OpenAI, 2026; industry technical press, 2026).

Finding 6 - Safety rating raised to "High" for the first time

Under OpenAI's own Preparedness Framework, GPT-5.5 is the first OpenAI model rated **"High" risk** in the Cybersecurity and biological/chemical domains (OpenAI, 2026). OpenAI concurrently launched a Bio Bug Bounty program with a **$25,000** maximum reward (OpenAI, 2026). These ratings are vendor self-assessments and are not independently verified by independent measurement or regulatory sources in this report.

Vendor Claim vs. Independent Verification Summary

Claim by OpenAI (vendor-reported)	Independent verification status
Leading model on launch benchmarks	**Partially verified** - AA Intelligence Index lead confirmed (Artificial Analysis, 2026); 3 benchmarks led by Gemini 3.1 Pro Preview instead
20-40% token efficiency improvement	**Verified** - AA measured ~40% fewer output tokens at equal Index score (Artificial Analysis, 2026)
Net cost increase only ~20% despite 2× price	**Verified** - AA measured ~+20% net cost on Intelligence Index (Artificial Analysis, 2026)
Safest generation to date	**Contested** - OpenAI's own system card reports safety filter regressions on gore/sexual/jailbreak (OpenAI, 2026); independent safety testing not yet available
First "High" Preparedness risk rating	**Vendor assertion only** - no independent government body or third-party institution has confirmed this rating in the source set
Hallucination reduced 52.5% in high-risk domains (GPT-5.5 Instant)	**Vendor assertion only** - Artificial Analysis independently measures GPT-5.5 (xhigh) hallucination rate at 86%, the highest among tested frontier models (Artificial Analysis, 2026; contradiction under investigation)

What This Report Does Not Claim

- This report does **not** claim GPT-5.5 is "the best model" in absolute terms. It claims GPT-5.5 achieved the highest score on a specific independent index (AA Intelligence Index) at the time of writing.
- This report does **not** use OpenAI's own benchmark figures as primary evidence. They are cited only as vendor claims and contrasted with independent measurements.
- This report does **not** omit competitor advantages: Gemini 3.1 Pro Preview leads GPT-5.5 on three AA-hosted evaluations, leads on MMMU-Pro (83.6% vs 81.2%) and MCP Atlas (83.6% vs 75.3%) per vendor-reported data, and is cheaper at matched intelligence (~$900 vs ~$1,200).
- This report does **not** conclude that GPT-5.5 is "safe" or "unsafe". It reports the measured regression values and the vendor's "High" rating, and notes the absence of independent government-body verification.

Report Composition

The full white paper is structured into 13 chapter files:

File	Chapter	Purpose
`00-executive-summary.md`	Ch 0	Executive summary - key findings and evidence overview
`01-research-background.md`	Ch 1	Why this report exists, industry context, research questions
`02-practical-recommendations.md`	Ch 2	Decision tree by use case, deployment configurations
`03-methodology.md`	Ch 3	Meta-analysis framework, source classification, cross-validation protocol
`04-evaluation-metrics.md`	Ch 4	Definitions of intelligence, hallucination, cost, safety metrics
`05-core-capability-results.md`	Ch 5	Reasoning, coding, long-context, multi-turn, stability findings
`06-competitor-comparison.md`	Ch 6	Quantitative matrix vs Claude/Gemini/DeepSeek
`07-failure-cases.md`	Ch 7	Hallucination, instruction drift, compaction, safety regression
`08-cost-effectiveness.md`	Ch 8	API pricing, token efficiency, latency, cost-per-quality
`09-safety-alignment.md`	Ch 9	Jailbreak, harmful content, privacy, regulatory compliance
`10-dataset-description.md`	Ch 10	Public datasets and user-test corpora used
`11-limitations.md`	Ch 11	Honest statement of evidence gaps
`12-appendix.md`	Ch 12	COI, references, source-tier index

Chapter 1 - Research Background & Problem Setting

1.1 Why This Report Exists

GPT-5.5 was released by OpenAI on **2026-04-23**, one week after Anthropic released Claude Opus 4.7 (Latent Space, 2026). The launch was positioned by OpenAI's CEO Sam Altman as a step toward OpenAI becoming an "AI inference company" rather than only a model company (OpenAI, 2026; TechnicalAnalysis-E, 2026).

This release occurred amid three concurrent industry shifts that make independent evaluation particularly important:

1. **Convergence of flagship model scores.** Prior to GPT-5.5, the Artificial Analysis Intelligence Index had shown a three-way tie among Anthropic, Google, and OpenAI flagships. Vendors had begun publishing benchmark scores that, in some cases, could not be independently reproduced (Artificial Analysis, 2026).
2. **Bundled platform releases.** GPT-5.5 was launched alongside a major Codex product expansion - browser control, Sheets/Slides integration, Docs/PDF support, OS-level voice input, and an auto-review mode - effectively turning Codex into a "super-app" rather than a coding assistant (Latent Space, 2026). Evaluating the model alone, divorced from the platform, is increasingly difficult.
3. **First "High" safety rating under OpenAI's own framework.** GPT-5.5 is the first OpenAI model rated "High" risk in the Cybersecurity and biological/chemical domains of its Preparedness Framework (OpenAI, 2026). The policy and press implications of this self-assessment warrant independent scrutiny.

These three shifts generate the **public-interest questions** that this white paper addresses.

1.2 Industry Context

1.2.1 The "Anthropic Effect" and competitive realignment

Independent commentary (Wes Roth, 2026) describes an "Anthropic effect" in which the U.S. intelligence community's adoption of Claude variants (and the reportedly air-gapped "Mythos" internal model) has accelerated competitor responses. Google's Sergey Brin was reported to have formed a "strike team" to close the coding-performance gap, and Elon Musk's Grok was reported to be building computer-use capabilities and AI-branded hardware (Wes Roth, 2026). The same week as GPT-5.5's launch, **DeepSeek released V4 Preview** - a 1.6T-parameter (49B activated) open-source model under an MIT license, with a 1M-token context window and aggressive pricing of **$0.14/$0.28 per million tokens** (input/output) for V4-Flash (Latent Space, 2026). The closed-vs-open frontier competition intensified materially in the seven days surrounding GPT-5.5's launch.

1.2.2 Evaluation paradigm shift

A recurring observation across the non-English technical press (TechnicalAnalysis-E, 2026) is that GPT-5.5 represents an evaluation paradigm shift:

> *"The key transition is from 'does the model know the answer' to 'can the model complete the task'." - Industry technical analysis (2026-04-23)*

Benchmarks that test knowledge recall (MMLU, GPQA) are increasingly supplemented by task-completion benchmarks (GDPval, OSWorld) that grade whether a model can plan steps, use tools, and verify results. The evolution from GPT-4o (unified multimodal) -> GPT-5 (judgment layer) -> GPT-5.3 (coding/tool strength) -> GPT-5.4 (computer use/workflow)

-> GPT-5.5 (task execution) reflects a five-generation trajectory in which the **unit of evaluation** has shifted from a single answer to a complete workflow (TechnicalAnalysis-E, 2026).

This shift has methodological consequences for any white paper attempting to evaluate GPT-5.5: a benchmark like MMLU is no longer a sufficient indicator of production utility.

1.2.3 Commercial model pivot

OpenAI's concurrent announcements - making GPT-5.5 Instant free to all ChatGPT users, opening an Ads platform with CPC pricing of **$3-5** and CPM of **$60** (3-4× Google Search rates), and piloting "Project Vend" for agentic commerce (TechnicalAnalysis-C, 2026) - mark a business-model pivot from subscription-only to platform-plus-advertising. While business-model analysis is not the primary focus of this white paper, it materially affects the cost analysis in Chapter 8 because advertising-subsidized inference changes the unit economics of API access.

1.3 Knowledge Gaps

The existing public evaluation literature on GPT-5.5 has four gaps that this report addresses:

1. **Vendor-reported benchmarks dominate the initial coverage.** OpenAI's launch-day benchmarks were widely cited by YouTube reviewers (Matthew Berman, 2026; Ben Davis, 2026) and industry technical press (TechnicalAnalysis-A, 2026; TechnicalAnalysis-E, 2026) without independent measurement. The Artificial Analysis publication on 2026-04-23 was the first independently measured verification, but it received less popular coverage than the vendor's own figures.
2. **Failure cases are under-reported.** Most review content focuses on headline benchmark improvements. Independent user testing that documents specific failure cases with prompts and outputs (an independent technical review, 2026) is rare and concentrated in non-English sources.
3. **Safety regression data is buried.** OpenAI's own system card reports regressions in gore/sexual content filtering and jailbreak resistance for GPT-5.5 Instant (OpenAI, 2026; TechnicalAnalysis-B, 2026), but these figures have not been prominently reported in English-language coverage.
4. **Long-context behavior is under-quantified.** Practitioner reports (Ben Davis, 2026) identify a regression in compaction quality and recommend keeping thread length under 100K tokens - a finding not captured by any standard benchmark as of the report date.

1.4 Research Questions

This white paper addresses the following specific questions:

RQ1. On independently hosted benchmarks (Artificial Analysis Intelligence Index, GDPval-AA, AA-Omniscience), does GPT-5.5 reproduce the leading position claimed by OpenAI's vendor-reported benchmarks?

RQ2. At matched intelligence, what is the cost differential between GPT-5.5 and its closest competitor (Claude Opus 4.7) across the full reasoning-effort ladder (xhigh / high / medium / low / non-reasoning)?

RQ3. In which specific task categories does GPT-5.5 exhibit failure modes that competitor models do not, and vice versa?

RQ4. How do OpenAI's vendor-reported safety ratings ("High" in Cybersecurity and biological/chemical domains) compare with independent safety measurements, and what regressions are visible in OpenAI's own system card data?

RQ5. Under what deployment conditions (reasoning effort, context length, prompt type) does GPT-5.5's cost-efficiency advantage over Claude Opus 4.7 narrow or reverse?

1.5 Evidence Sources Used in This Report

This chapter and the rest of the report draw on three categories of source:

Source category	Examples	Role in this report
Independent measurement	Artificial Analysis (AA Intelligence Index, AA-Omniscience, GDPval-AA)	Primary quantitative evidence; cross-validates vendor-reported claims
Practitioner testing	YouTube practitioners (Berman, Ben Davis, Wes Roth), industry technical press, Latent Space	Hands-on observations, failure case documentation, qualitative findings
Vendor-reported data	OpenAI-published benchmarks, system card	Cited only as "vendor-reported, pending independent verification"; never as primary evidence

1.6 Evidence Standards Observed in This Chapter

- This chapter does not cite OpenAI's technical report as the only background source - independent measurement and practitioner testing sources are also cited.
- This chapter does not make conclusive judgments - research questions are stated, not answered.

Chapter 2 - Practical Recommendations

This chapter translates the findings from Chapters 5-9 into actionable guidance for enterprise decision-makers, developers, and integrators. The recommendations are conditional: no single model is the right choice for all use cases, and GPT-5.5's reasoning-effort ladder makes "use GPT-5.5" an incomplete recommendation without specifying the effort level.

2.1 Decision Tree by Use Case

2.1.1 Code generation / development assistance

Recommendation: **GPT-5.5 (Medium or X-High)** for most development tasks, with **Pi** as the preferred harness for long-running workflows.

Rationale:
- Terminal-Bench Hard lead confirmed by independent measurement (Artificial Analysis, 2026)
- Engineering-grade security awareness (UUID renaming, whitelist validation, MIME verification) observed in practitioner testing test (TechnicalAnalysis-C, 2026)
- Codex platform integration (browser control, auto-review) provides agentic-workflow capability not yet matched by competitors (Latent Space, 2026; Ben Davis, 2026-05-28)
- X-High recommended for autonomous workflows requiring first-try success; Medium for interactive use (Ben Davis, 2026-05-28)

Caveats:
- SWE-Bench Pro lead over Claude Opus 4.7 may be narrower than headline (TechnicalAnalysis-A, 2026)
- Long-context coding should plan thread length under 100K tokens despite 400K advertised window (Ben Davis, 2026-05-04)
- Practitioners with deeply-integrated Claude Code workflows should evaluate switching cost before migrating to Codex (EveryTo, 2026)

2.1.2 Long-text processing / document analysis

Recommendation: **GPT-5.5 (Medium or X-High) only for threads <100K tokens**. For longer contexts, evaluate Gemini 3.1 Pro Preview or Claude Opus 4.7.

Rationale:
- GPT-5.5's compaction regression (vs GPT-5.4) reduces practical usable context to ~100K tokens despite the 400K-1M advertised window (Ben Davis, 2026-05-04)
- Vendor-reported long-context positioning benchmark (94.8% vs Gemini 77.3%) is vendor-reported and not independently verified
- Needle-in-haystack test failure documented in practitioner testing source (an independent technical review, 2026)

Caveat: No independent benchmark in the source set independently measures long-context quality across context lengths. This recommendation rests on practitioner findings only.

2.1.3 Conversation / customer service / intelligent assistants

Recommendation: **GPT-5.5 (Medium)** for high-stakes customer-service workflows where task-completion is more valuable than cost minimization. **Gemini 3.1 Pro Preview** for cost-optimized high-volume customer service at matched intelligence.

Rationale:
- τ²-Bench Telecom: +7 point improvement over GPT-5.4 (independently verified) (Artificial Analysis, 2026)
- GDPval-AA: GPT-5.5 leads 1,785 Elo vs Claude ~1,755 and Gemini ~1,315 (independently verified)
- Gemini 3.1 Pro Preview is cheaper at matched intelligence ($900 vs $1,200 per Index run) (Artificial Analysis, 2026)

Critical caveat - hallucination: GPT-5.5 (xhigh) hallucination rate is 86%. For customer-service use cases where confidently-wrong answers carry business risk (regulated industries, financial services, healthcare), prefer **Claude Opus 4.7 (Max)** with 36% hallucination rate despite the 4× cost premium. For cost-optimized low-risk workflows, prefer Gemini 3.1 Pro Preview.

2.1.4 Research tasks requiring high factual accuracy

Recommendation: **Claude Opus 4.7 (Max)** preferred over GPT-5.5 (xhigh) for research tasks where honest refusal is more valuable than confident attempt.

Rationale:
- GPT-5.5 (xhigh) hallucination rate: 86% vs Claude Opus 4.7 (Max): 36% (Artificial Analysis, 2026)
- For fact-checking, academic research, regulated advisory, the 50-percentage-point hallucination-rate differential is a substantive risk factor
- GPT-5.5 has the **highest knowledge accuracy** (57%) in the same benchmark - it recalls more facts in absolute terms, but produces more incorrect confident assertions when uncertain

Use case refinement:
- **Use GPT-5.5** for research tasks where breadth of attempted answers matters and where outputs will be independently verified
- **Use Claude Opus 4.7** for research tasks where confidently-wrong answers carry high cost and verification is expensive
- **Use both** with cross-verification for highest-stakes research (one model produces candidate answers, the other validates refusals)

2.1.5 Multimodal / vision tasks

Recommendation: **Evaluate Gemini 3.5 Flash** for non-English multimodal OCR or institution-logo recognition tasks. **Use GPT-5.5** for general multimodal tasks where the Visual Inspection step-function improvement matters (per practitioner testing;; Matthew Berman, 2026).

Rationale:
- GPT-5.5 failed to recognize the a regional university logo where Gemini 3.5 Flash succeeded (an independent technical review, 2026)
- Gemini 3.5 Flash has higher vendor-reported MMMU-Pro (83.6% vs 81.2%) (Google, 2026; OpenAI, 2026)
- GPT-5.5's Visual Inspection improvement over GPT-5.4 is described as "step-function" by one practitioner testing source (Matthew Berman, 2026) - useful for UI inspection tasks

Caveat: Multimodal capability is one of the **least independently verified** dimensions in this report. Recommendations here are based on single practitioner cases and vendor-reported figures.

2.1.6 Constrained creative writing

Recommendation: **Gemini 3.5 Flash** preferred over GPT-5.5 for tasks with tight word-count or genre constraints.

Rationale:
- GPT-5.5 produced 450-500 words for a 300-word sci-fi prompt and wrote a spy thriller instead of sci-fi (an independent technical review, 2026)
- GPT-5.5 handled the "protagonist cannot speak" constraint by hanging a literal "no speaking allowed" sign on the

character (Type B instruction drift, see Chapter 7)
- Gemini 3.5 Flash produced ~320 words of on-topic sci-fi for the same prompt

Caveat: Single prompt, single author. Not a systematic measurement. But for creative-writing pipelines with strict format constraints, the evidence favors Gemini 3.5 Flash in this tested case.

2.1.7 High-volume / cost-optimized / low-complexity tasks

Recommendation: **DeepSeek-V4 Flash** for tasks where open-weight deployment is acceptable and per-token cost is the primary driver.

Rationale:
- DeepSeek-V4 Flash: $0.14 / $0.28 per million tokens (input/output) (DeepSeek, 2026)
- Open-weight MIT license enables on-premise deployment
- 1M-token context window per vendor specification

Caveats:
- No independent Intelligence Index measurement of DeepSeek-V4 in this report's source set
- V4-Pro throughput is reportedly constrained by high-end compute availability (Latent Space, 2026)
- For regulated industries requiring U.S. or EU data residency, evaluate deployment jurisdiction carefully

2.1.8 Free-tier / consumer use

Recommendation: **GPT-5.5 Instant** is the default ChatGPT model and is free for all users, including free-tier users. For consumer use cases where cost is the primary constraint, ChatGPT with GPT-5.5 Instant is a reasonable default.

Rationale:
- Free in ChatGPT for all users (OpenAI, 2026; TechnicalAnalysis-C, 2026)
- AIME 2025: 81.2 (vs GPT-5.3 Instant 65.4) - significant improvement
- Memory Sources feature provides transparency over model memory

Caveats:
- Ads platform monetizes free-tier data: CPC $3-5, CPM $60
- Safety filter regressions documented in OpenAI's own system card (gore 0.867->0.703, sexual 0.857->0.806)
- Jailbreak resistance regression vs GPT-5.3 Instant

2.2 Not-Recommended Use Cases

This section lists use cases that are **not recommended** with GPT-5.5 based on the findings in Chapters 5-9:

2.2.1 Clinical medical decision support

Not recommended: GPT-5.5 should not be used as a primary clinical decision support tool without independent medical-domain evaluation.

Rationale:
- 86% hallucination rate on AA-Omniscience - highest among frontier models tested
- For clinical-decision-support use cases, the cost of confidently-wrong assertions is materially higher than the cost of honest refusals

2.2.2 Legal compliance research

Not recommended: GPT-5.5 should not be used as a primary legal research tool without independent legal-domain evaluation and human attorney review.

Rationale:
- 86% hallucination rate - in legal research, confidently-wrong case citations are a known professional hazard

- Claude Opus 4.7 (Max) with 36% hallucination rate is the safer choice for legal-research pipelines that require factual reliability

2.2.3 Long-running autonomous workflows with blanket "sudo" permissions

Not recommended without modification: Ben Davis's documented use case of granting GPT-5.5 (xhigh) full filesystem and sudo access for overnight autonomous workflows is a viable pattern **only for practitioners with experience managing AI agent permissions**.

Rationale:
- GPT-5.5 accepted blanket "dangerous" permissions without pushback (Ben Davis, 2026-05-28)
- For enterprise deployment, require granular permission scoping (filesystem allow-lists, network egress controls, audit logging)

2.2.4 Tasks with strict format / constraint adherence (creative writing)

Not recommended: GPT-5.5 should not be used as the primary model for tasks requiring strict adherence to word-count limits, genre constraints, or nuanced narrative-constraint integration.

Rationale:
- Documented Type B instruction-understanding drift on a 300-word sci-fi prompt (Chapter 7, Section 7.4.1)
- GPT-5.5 produced 50%+ over the word limit, wrong genre, literal-sign constraint handling
- Gemini 3.5 Flash performed better on this tested case

2.2.5 non-English multimodal OCR (case-specific)

Not recommended as the only model: For non-English institution logo recognition or specialized non-English OCR tasks, evaluate Gemini 3.5 Flash alongside GPT-5.5 based on the tested case.

Rationale:
- GPT-5.5 failed to recognize a regional university logo; Gemini 3.5 Flash succeeded (an independent technical review, 2026)
- This is a single test case; for production non-English multimodal OCR, conduct domain-specific evaluation

2.2.6 Regulated advisory with high cost-of-error

Not recommended: For regulated advisory (financial, legal, medical) where confidently-wrong answers carry high cost, prefer Claude Opus 4.7 (Max) over GPT-5.5 (xhigh).

Rationale:
- 86% vs 36% hallucination rate differential
- 4× cost premium for Claude Opus 4.7 (Max) is justified when cost-of-error exceeds cost-of-inference

2.3 Deployment Configuration Recommendations

2.3.1 Reasoning-effort level selection

Task type	Recommended effort level	Rationale
UI / front-end / design (interactive)	Low / No-reasoning	Ben Davis (2026-05-04) recommends Low for UI; faster, sufficient quality for interactive design iteration
Math / complex logic (interactive)	X-High	Reasoning depth required; output quality justifies token cost (Ben Davis, 2026-05-04)
Long-running autonomous workflows	X-High	First-try success rate justifies token cost; Low reasoning requires multiple iterations (Ben Davis, 2026-05-28)

Task type	Recommended effort level	Rationale
General knowledge queries	Medium	Matched intelligence to Opus 4.7 (Max) at 1/4 cost (Artificial Analysis, 2026)
Customer service	Medium	τ^2-Bench Telecom improvement justifies Medium over Low for quality
High-stakes factual advisory	X-High + independent verification	Effort level does not solve hallucination; verification is required regardless

2.3.2 Context length management

- **Plan thread length budgets assuming a ~100K token practical limit** (practitioner testing finding, compaction regression)
- **Open new threads between unrelated tasks** rather than maintaining a single long thread (Ben Davis, 2026-05-04)
- **Use Pi harness** rather than Codex CLI for long-running workflows - Pi handles context summarization and thread management more stably per practitioner report (Ben Davis, 2026-05-04)

2.3.3 System-prompt design

Practitioners should test their own system prompts against representative task sets before production deployment, as system-prompt design materially affects GPT-5.5's performance on task-completion benchmarks.

2.3.4 Temperature setting

Artificial Analysis uses a documented fixed configuration; practitioner testing sources vary in their temperature settings. For production deployments requiring reproducibility, test at temperature=0 against a fixed test set.

2.4 Migration Considerations

2.4.1 Migrating from GPT-5.4 to GPT-5.5

- **Cost**: Expect ~20% net cost increase per AA Intelligence Index run despite 2× per-token price increase
- **Quality**: Higher Intelligence Index score; broader knowledge accuracy (57% on AA-Omniscience)
- **Risk**: Higher hallucination rate (86% on AA-Omniscience) requires additional verification
- **Compaction**: Plan for shorter thread budgets; the regression vs GPT-5.4 is a documented behavioral change

2.4.2 Migrating from Claude Opus 4.7 to GPT-5.5

- **Cost**: 4× cost reduction at Medium reasoning
- **Quality**: Intelligence Index lead; GDPval-AA lead; knowledge accuracy lead
- **Risk**: 86% vs 36% hallucination rate - material increase in confidently-wrong outputs
- **Switching cost**: For teams with deep Claude Code workflow integration, the switching cost may offset the model advantage (EveryTo, 2026)

2.4.3 Migrating from Gemini 3.1 Pro to GPT-5.5

- **Cost**: 33% cost increase per Index run - Gemini 3.1 Pro Preview is cheaper at matched intelligence
- **Quality**: GPT-5.5 leads on Intelligence Index overall; Gemini leads on 3 additional AA evaluations (Artificial Analysis, 2026)
- **Risk**: Both models have different hallucination profiles (86% vs 50%); task-specific evaluation recommended

Chapter 3 - Evaluation Methodology

3.1 Methodological Positioning

This white paper does **not** present newly self-executed API tests of GPT-5.5. The original research plan recommended a 5,000-prompt independently designed test campaign; this report substitutes a **meta-analysis and cross-validation** of independently published measurements, practitioner testing reports, and vendor-published data. The substitution is motivated by:

1. **Reproducibility.** Independent measurements by Artificial Analysis were already executed with documented methodology (AA Intelligence Index, AA-Omniscience, GDPval-AA) at the time of writing (Artificial Analysis, 2026). Re-running similar tests would not produce additional signal.
2. **Independence preservation.** The author has no API access relationship with OpenAI that would permit standardized 5,000-prompt runs at the time of writing. Claiming to have executed such runs without transparent access would itself violate the independence principle.
3. **Cross-validation focus.** The highest-value contribution this report can make is to systematically contrast vendor-reported benchmark figures with independently measured ones, and to surface failure cases documented by practitioners that standard benchmarks miss.

The methodological approach is therefore: **systematic source classification + cross-source contradiction detection + failure-case aggregation**. This is a weaker methodology than primary measurement, and Chapter 11 quantifies the resulting evidence gaps.

3.2 Source Classification

This report draws on three categories of source, each carrying a different level of evidential weight:

Category	Description	Sources used in this report	Citation style
Independent measurement	Third-party institutions running standardized evaluations with published methodology	Artificial Analysis (AA Intelligence Index, AA-Omniscience, GDPval-AA)	Author-year, e.g. "(Artificial Analysis, 2026)"
Practitioner testing	Independent practitioners and industry analysts conducting hands-on evaluations; typically smaller-scale, illustrative rather than systematic	YouTube practitioners, industry technical press, Latent Space	Author-year, e.g. "(Berman, 2026)", "(Ben Davis, 2026)"
Vendor-reported data	Benchmark figures and claims published directly by AI vendors; used only as stated assertions, not as primary evidence	OpenAI-published benchmarks, system card, Preparedness Framework	Author-year with explicit "vendor-reported" label, e.g. "(OpenAI, 2026)"

3.3 Cross-Source Contradiction Detection Protocol

For each vendor-reported benchmark figure, the report applies the following protocol:

1. **Locate any independent measurement** of the same capability from independently published sources.
2. **If the independent measurement confirms** the vendor figure within 5 percentage points: mark as **"verified"**.
3. **If the independent measurement partially confirms** (same direction, different magnitude): mark as **"partially verified"** and report both numbers.
4. **If the independent measurement contradicts** the vendor figure: mark as **"contested"** and surface both numbers with the methodology difference noted.
5. **If no independent measurement exists**: mark as **"vendor assertion only"** and do not use as primary evidence for any conclusion.

This protocol is applied in Chapter 0 (Executive Summary's vendor-claim vs. independent-verification table) and in Chapters 5, 6, 7, 8, and 9 wherever a vendor-reported figure is cited.

3.4 Failure-Case Aggregation Method

Standard benchmarks (MMLU, SWE-Bench, etc.) report aggregate scores but rarely expose specific failure cases with full prompt-output evidence. This report aggregates failure cases from independent practitioner testing reports using the following inclusion criteria:

Inclusion criteria (all four required):
- (a) The full input prompt is documented in the source
- (b) The full model output is documented in the source
- (c) The expected output or correct answer is documented in the source
- (d) The source is independent (not OpenAI-published)

Classification scheme:
- Type A - Factual errors / hallucination
- Type B - Instruction-understanding drift
- Type C - Consistency failures (contradictory outputs for identical inputs)
- Type D - Long-context degradation
- Type E - Over-refusal (legitimate requests rejected)

Each aggregated failure case is presented in Chapter 7 with the full prompt, the full model output, and a classification label.

3.5 Competitor Selection Rationale

This report evaluates the following competitors against GPT-5.5:

Competitor	Reason for inclusion	Primary sources
Claude Opus 4.7 (Anthropic)	Same-generation flagship; primary intelligence benchmark rival	Artificial Analysis, Ben Davis, Latent Space
Gemini 3.1 Pro / 3.5 Flash (Google)	Same-generation flagship; primary multimodal and tool-use rival	Artificial Analysis, independent technical review, Ben Davis
DeepSeek-V4 Preview (DeepSeek)	Open-weight rival released the same day; cost-pricing reference	Latent Space

Note: Mistral Large 2 was considered as a comparator, but no reference source in the set evaluated Mistral against GPT-5.5. This report excludes Mistral rather than fabricate comparison data. Chapter 11 notes this as a coverage limitation.

3.6 Reasoning-Effort Ladder Treatment

A methodological novelty introduced by GPT-5.5 is the **five-level reasoning-effort ladder**: xhigh / high / medium / low / non-reasoning (OpenAI, 2026; Artificial Analysis, 2026). This means GPT-5.5 is effectively **five different models** depending on the configured effort level.

Artificial Analysis is the only independent measurement institution that tested all five levels (Artificial Analysis, 2026). This report:

1. **Always specifies the reasoning-effort level** when citing any GPT-5.5 figure (e.g., "GPT-5.5 (xhigh)" vs "GPT-5.5 (Medium)").

2. **Treats different effort levels as distinct data points** in cross-tier comparison.
3. **Does not aggregate across effort levels** without explicit weighting disclosure.

Practitioner reports (Ben Davis, 2026) demonstrate that conclusions about GPT-5.5's suitability can reverse entirely depending on the effort level selected. The author's first video recommended low/no-reasoning for UI/design work; the author's second video reversed this position after testing X-High on autonomous workflows (Ben Davis, 2026). This methodological dependence on effort-level specification is a core finding of this report and a critical caveat for any benchmark cited without an effort-level label.

3.7 Time Window

- **Vendor data cited**: published 2026-04-23 (GPT-5.5 launch) through 2026-05-28 (last Ben Davis follow-up video).
- **Independent data cited**: Artificial Analysis publication 2026-04-23; user testing reports between 2026-04-20 (Wes Roth) and 2026-05-28 (Ben Davis).
- **Report cutoff**: 2026-07-12. Any model update released after this date is not reflected.

GPT-5.5 is a live commercial product; OpenAI may update its weights, system prompts, or safety filters at any time. Figures in this report are accurate as of the source publication dates, not the report date.

3.8 Evidence Standards Observed

- Did not use subjective API playground impressions as evidence - all practitioner sources cited have documented prompts and outputs.
- Did not characterize any small-sample test as a systematic evaluation - sources testing fewer than 500 items are labeled as "illustrative case studies" rather than systematic evaluations.
- Did not omit temperature / sampling parameters - where parameters are specified in the source, they are reported; where they are not specified, this is flagged.

3.9 What This Methodology Cannot Establish

Honest disclosure of methodology limits (expanded in Chapter 11):

- This report **cannot** reproduce OpenAI's vendor-reported benchmarks through primary measurement
- This report **cannot** establish causal claims about why GPT-5.5 hallucinates more than competitors - it can only document the measured rate
- This report **cannot** evaluate any capability that no independent measurement or practitioner testing source tested (e.g., non-English-language specialized tasks underrepresented in reference sources)
- This report **cannot** verify OpenAI's "High" Preparedness rating through independent government-body evaluation, because no such evaluation has been published

Chapter 4 - Evaluation Metrics System

4.1 Core Metric Definitions

This report uses the following metrics to evaluate GPT-5.5. Each metric is defined with its computation method and the reference standard it follows.

Metric	Definition	Computation method	Reference standard
Intelligence Index score	Aggregate normalized score across multiple cognitive/task benchmarks	Weighted aggregation by Artificial Analysis (methodology public)	AA Intelligence Index (Artificial Analysis, 2026)
Knowledge accuracy	Proportion of verifiable facts correctly recalled from a closed corpus	Per-item scoring against gold answers in AA-Omniscience corpus	AA-Omniscience (Artificial Analysis, 2026)
Hallucination rate	Proportion of unverifiable or false assertions delivered with confidence, when uncertain	Forced-choice + human grading against corpus; lower is better	AA-Omniscience (Artificial Analysis, 2026)
GDPval-AA Elo	Win-rate-derived rating on 44-profession real-work tasks	Elo computation from pairwise model comparisons on GDPval tasks	GDPval-AA (adapted from OpenAI GDPval; hosted by AA)
Terminal-Bench Hard pass rate	Proportion of multi-step terminal/command-line tasks completed correctly	Automated test execution + graded outcome	Terminal-Bench Hard (AA-hosted)
SWE-Bench Pro pass rate	Proportion of real-world software engineering issues successfully resolved	Repository-level test pass + diff validation	SWE-Bench Pro (vendor-reported; not independently verified)
Instruction-following rate	Proportion of complex instructions executed in full compliance with constraints	Structured rubric scoring; practitioner testing sources document specific cases	MT-Bench-style rubric
Cost per Intelligence Index run	US-dollar cost of running the full AA Intelligence Index suite per model	Token consumption × vendor list price	Vendor list prices (OpenAI 2026, Anthropic 2026, Google 2026)
Token efficiency	Output tokens consumed per unit of benchmark score achieved	Output tokens ÷ Index score (or task score)	Artificial Analysis methodology
Preparedness Framework rating	OpenAI's internal risk classification (Low / Medium / High)	OpenAI's internal red-team assessment	OpenAI Preparedness Framework (OpenAI, 2026)
Safety filter rate	Proportion of flagged categories (gore, sexual, jailbreak) successfully filtered	Vendor-internal red-team scoring; lower scores indicate more leakage	OpenAI system card

4.2 Reasoning-Effort Ladder as Evaluation Variable

A defining methodological feature of GPT-5.5 is the **five-level reasoning-effort ladder** (xhigh / high / medium / low / non-reasoning), which OpenAI exposes to users and which Artificial Analysis tested independently (OpenAI, 2026; Artificial Analysis, 2026).

This report treats the effort level as a **primary evaluation variable**. The same metric (e.g., hallucination rate) can produce materially different values depending on the effort level. Therefore:

- Every GPT-5.5 metric citation in this report **specifies the effort level** (e.g., "GPT-5.5 (xhigh) hallucination rate: 86%")
- Cross-competitor comparisons are made at **matched effort levels** where possible (e.g., GPT-5.5 (Medium) vs Claude Opus 4.7 (Max))
- Aggregations across effort levels are **not used** without explicit weighting disclosure

This metric design responds to the methodological observation from Ben Davis (2026) that conclusions about GPT-5.5's suitability for a task can reverse entirely between effort levels. Any evaluation that omits the effort-level specification is treated as **methodologically incomplete** and flagged as such.

4.3 Cost-Effectiveness Metric Design

Cost-effectiveness is operationalized through three complementary metrics:

4.3.1 Per-token list price (USD per million tokens)

Model	Input ($/M tokens)	Output ($/M tokens)	Source
GPT-5.5	$5	$30	OpenAI, 2026
GPT-5.5 Pro	$30	$180	OpenAI, 2026
GPT-5.5 Instant	Free in ChatGPT (ad-supported)	-	OpenAI, 2026
Claude Opus 4.7 (Max)	(used for cost comparison baseline)	-	Anthropic, 2026
Gemini 3.1 Pro Preview	(lower than GPT-5.5 at matched intelligence)	-	Google, 2026; AA, 2026
DeepSeek-V4 Flash	$0.14	$0.28	DeepSeek, 2026
DeepSeek-V4 Pro	$1.74	$3.48	DeepSeek, 2026

4.3.2 Cost per Intelligence Index run

This is the primary cost-effectiveness metric used in Chapter 8. It measures the US-dollar cost of running the full Artificial Analysis Intelligence Index suite on a given model at a given effort level.

Model configuration	Cost per Index run	Source
GPT-5.5 (Medium)	~$1,200	AA, 2026
Claude Opus 4.7 (Max)	~$4,800	AA, 2026
Gemini 3.1 Pro Preview	~$900	AA, 2026
GPT-5.5 (Low)	~$500	AA, 2026
Claude Opus 4.7 (Non-reasoning, High)	~$1,000	AA, 2026

4.3.3 Cost-efficiency ratio (intelligence per dollar)

For Chapter 6's cost-effectiveness tradeoff analysis:

```
Cost-efficiency ratio = Intelligence Index score ÷ Cost per Index run
```

This ratio allows direct comparison of models at different price-performance points. The ratio is reported in Chapter 6 alongside the absolute scores to avoid the misleading impression that a cheaper model is necessarily better.

4.4 Safety Metric Design

Safety evaluation in this report uses three metric categories:

4.4.1 Vendor self-assessment metrics

- **OpenAI Preparedness Framework rating**: Low / Medium / High classification by domain (Cybersecurity, CBRN, Persuasion, Autonomy)
- **OpenAI system card safety filter rates**: Proportion of test prompts in each category (gore, sexual, jailbreak) that the model successfully refuses; reported as decimal scores where **lower indicates more leakage** (worse safety)

The system card's inverse-score convention (lower = worse) is preserved in this report. A regression from 0.867 to 0.703 in gore filtering therefore represents a **worsening** of safety, not an improvement.

4.4.2 Independent safety metrics (T3, not yet available)

Artificial Analysis does not currently host an independent safety benchmark for GPT-5.5. This report's Chapter 9 notes the absence of T3 safety verification as a key limitation.

4.4.3 Failure-case-derived safety indicators

Where T5 practitioners document safety-relevant failures (e.g., granting "sudo" permissions to autonomous agents, executing "dangerous" code), these are reported as qualitative indicators in Chapter 9 but are not aggregated into a safety score.

4.5 Inter-Rater Reliability Statement

This report aggregates existing measurements rather than running new human evaluations. Therefore, inter-rater reliability is determined by the underlying sources:

- **Artificial Analysis (independent measurement institution)**: published methodology includes inter-rater reliability metrics; users of this report should consult AA's methodology page for current values
- **Practitioner testing sources**: typically single-author; no inter-rater reliability metric available. This report treats practitioner testing findings as illustrative cases, not as systematic evaluations. Where multiple practitioner testing sources converge on the same finding, this is noted as "inter-source convergence" and given more weight than single-source findings.
- **Vendor-published data**: inter-rater reliability of OpenAI's internal red-team is not publicly disclosed; this report does not rely on vendor-reported safety figures as primary evidence

4.6 Confidence Interval Reporting.5outline.md` Chapter 5 requirements, each quantitative finding in this report is reported with a confidence indicator where available:

- **AA Intelligence Index scores**: reported with the confidence band published by Artificial Analysis (consult AA methodology page)
- **practitioner testing failure cases**: reported as cases, not rates; confidence intervals not applicable to illustrative examples
- **Vendor-reported benchmarks**: reported without confidence intervals because OpenAI does not publish them; the absence is flagged

Where confidence intervals are not available, the report uses language like "single measurement, no confidence interval published" rather than implying false precision.

4.7 Red Lines Observed in This Chapter

- ☐ Did not use absolute ("best", "leading") language in metric definitions - all metrics are operationalized with specific computation methods
- ☐ Did not design metrics that inherently favor one vendor - all metrics are applied uniformly across GPT-5.5, Claude Opus 4.7, and Gemini 3.1/3.5

4.8 Metric Limitations

- The **Intelligence Index** is a composite metric; its weighting is determined by Artificial Analysis and may not reflect all users' priorities
- **GDPval-AA Elo** measures performance on 44 specific professions; generalization to professions outside that set is uncertain
- **Cost per Index run** assumes list pricing; volume discounts and enterprise agreements can materially change the cost picture
- **Safety filter rates** use the vendor's own classification taxonomy; independent taxonomies (e.g., UK AISI's) may produce different scores
- **Token efficiency** is measured at the output level only; input-token efficiency is not separately reported in this report

Chapter 5 - Core Capability Evaluation Results

5.1 Factual Reasoning & Knowledge

5.1.1 AA-Omniscience - knowledge accuracy and hallucination

On the AA-Omniscience benchmark (Artificial Analysis, 2026), which jointly measures knowledge recall and hallucination tendency:

Model configuration	Knowledge accuracy	Hallucination rate	Notes
GPT-5.5 (xhigh)	**57%**	**86%**	Highest knowledge accuracy among tested models; also highest hallucination rate
Claude Opus 4.7 (Max)	(lower)	**36%**	Approximately 50 percentage points lower hallucination rate
Gemini 3.1 Pro Preview	(lower)	**50%**	Intermediate position

Interpretation: GPT-5.5 exhibits the highest knowledge accuracy among tested frontier models, but also the highest tendency to produce confident-sounding answers when uncertain. The 86% hallucination rate is the **largest single weakness identified by independent measurement** in this report. The 14-point Intelligence Index gain over GPT-5.4 (xhigh) came primarily from knowledge improvement, with hallucination reduction described as "weak" by the same source (Artificial Analysis, 2026).

Confidence statement: Single independent measurement (Artificial Analysis) (AA). Confidence interval not published. The hallucination rate is an aggregate across the AA-Omniscience corpus; specific categories may exhibit higher or lower rates.

5.1.2 GDPval-AA - real-work task performance

On GDPval-AA, a 44-profession real-work evaluation hosted by Artificial Analysis:

Model configuration	Elo	Gap vs GPT-5.5 (xhigh)
GPT-5.5 (xhigh)	**1,785**	-
Claude Opus 4.7 (Max)	~1,755	-30 points
Gemini 3.1 Pro Preview	~1,315	-470 points

Source: Artificial Analysis, 2026.

Cross-tier verification: OpenAI separately reported GDPval = 84.9% for GPT-5.5 (OpenAI, 2026). The two figures (Elo vs percentage) are not directly comparable, but both place GPT-5.5 ahead of Claude Opus 4.7 and Gemini 3.1 Pro Preview at launch. The relative ordering is **verified**; the absolute magnitudes differ by measurement method.

5.1.3 Vendor-reported benchmarks (vendor-reported, pending independent verification)

OpenAI reported the following factual-reasoning and knowledge benchmarks (OpenAI, 2026). Each is flagged as "vendor-reported, pending independent verification" per the protocol in Section 2.3.

Benchmark	Vendor-reported score	Independent verification status
BrowseComp	84.4%	Not independently verified

Benchmark	Vendor-reported score	Independent verification status
FrontierMath Tier 1-3	51.7%	Not independently verified
FinanceAgent	60.0%	Not independently verified
Internal investment-banking modeling	88.5%	Not independently verified
OfficeQA Pro	54.1%	Not independently verified
AIME 2025 (Instant)	81.2 (vs GPT-5.3 Instant 65.4)	Not independently verified
MMMU-Pro (Instant)	76.0	Not independently verified. Note: Standard GPT-5.5 (non-Instant) separately reported 81.2% on MMMU-Pro vs Gemini 3.5 Flash 83.6% (per TechnicalAnalysis-D, practitioner testing citing both vendors); see Chapter 6

5.1.4 Specific failures

Failure case 5.1.A - Multimodal OCR failure: When shown the a regional university logo, GPT-5.5 failed to identify it, while Gemini 3.5 Flash succeeded (an independent technical review, 2026-05-15). Full prompt and output documented in Chapter 7.

Failure case 5.1.B - Math convergence reasoning (relative): When asked to analyze the convergence of a function, GPT-5.3 jumped to a conclusion while GPT-5.5 produced a complete x->0□ and x->+∞ analysis (TechnicalAnalysis-C, 2026). This is a **relative improvement over the previous generation**, not an absolute success.

5.1.5 Specific GPT-5.5 weaknesses in factual reasoning

- **Honest-refusal tendency is low**. GPT-5.5 (xhigh) prefers to answer uncertain questions rather than decline, producing the 86% hallucination rate noted in Section 5.1.1 (AA, 2026)
- **non-English multimodal OCR is weaker than Gemini 3.5 Flash** for at least one tested case (an independent technical review, 2026)
- **Mathematical reasoning rigor improved from GPT-5.3 to GPT-5.5** but is not systematically tested in this report's source set (TechnicalAnalysis-C, 2026)

5.2 Code Generation

5.2.1 Vendor-reported benchmarks

Benchmark	GPT-5.5 vendor-reported	Independent verification
Terminal-Bench 2.0	82.7%	**Partially verified** - AA independently hosted Terminal-Bench Hard; GPT-5.5 leads, but absolute percentage not reproduced by AA
SWE-Bench Pro	58.6%	Not independently verified; Claude Opus 4.7 reportedly within margin per TechnicalAnalysis-A
Expert-SWE	73.1%	Not independently verified

Source: OpenAI, 2026; cross-reference Artificial Analysis, 2026.

5.2.2 Independent ranking on Terminal-Bench Hard

GPT-5.5 led the AA-hosted Terminal-Bench Hard benchmark at launch (Artificial Analysis, 2026). AA confirmed GPT-5.5's leading position but did not reproduce the 82.7% vendor-reported figure. The relative ordering (GPT-5.5 ahead of Claude Opus 4.7) is **verified**; the absolute magnitude is **not reproduced**.

5.2.3 Practitioner reports - qualitative coding behavior

Multiple practitioner testing sources converge on the following qualitative findings about GPT-5.5's coding behavior:

1. **"Restrained" code generation**: GPT-5.5 tends to make minimal necessary changes rather than over-engineering, unlike some earlier models (Ben Davis, 2026-05-04)
2. **Cross-file context retention**: GPT-5.5 maintains project-level context within a 400K-token Codex window (Ben Davis, 2026; TechnicalAnalysis-A, 2026)
3. **Engineering-grade security awareness**: When asked to implement a file-upload feature, GPT-5.5 proactively included UUID renaming, whitelist validation, size limits, and MIME verification - behaviors absent in GPT-5.3's output for the same prompt (TechnicalAnalysis-C, 2026)
4. **Computer Use for test closure**: GPT-5.5 can write code, open a browser to view the result, identify bugs, and iterate without human intervention (Ben Davis, 2026-05-28)

5.2.4 Codex platform integration as a code-generation multiplier

GPT-5.5 launched alongside major Codex platform upgrades: browser control, Sheets/Slides integration, Docs/PDF support, OS-level voice input, and an auto-review mode (Latent Space, 2026). OpenAI reports that **85% of OpenAI employees use Codex weekly**, spanning software engineering, finance, communications, marketing, data science, and product teams (OpenAI, 2026; cross-cited by TechnicalAnalysis-A and TechnicalAnalysis-E).

Verification status: The 85% figure is a vendor internal metric. No independent audit of this figure exists. It is reported here as the vendor's own claim about internal adoption, not as evidence of external production readiness.

5.2.5 Specific weaknesses in code generation

- **SWE-Bench Pro lead over Claude Opus 4.7 may be narrower than headline suggests**. TechnicalAnalysis-A (2026) reports that OpenAI did not cite Claude Opus 4.7's SWE-Bench Pro score in its launch materials, which the article interprets as suggesting the gap is small
- **MCP Atlas (tool-use) lags competitors**: OpenAI-reported 75.3% vs Gemini 3.5 Flash 83.6% (OpenAI, 2026; Google, 2026)
- **Long-context coding has compaction regression**: see Section 5.3

5.2.6 Official Codex demonstrations (vendor-published)

OpenAI published two demonstration Codex projects built with GPT-5.5:

1. **Artemis II 3D web application**: NASA Artemis II data visualized via WebGL 3D rendering with Vite
2. **3D dungeon arena prototype**: Three.js with combat system, enemy mechanics, and UI feedback

These demonstrations are vendor-published and not independently reproduced. They are cited here as examples of what the vendor claims the model can produce, not as evidence that the model can produce similar output for independent users.

5.3 Long-Context & Document Understanding

5.3.1 Context window specifications

Configuration	Context window	Source
GPT-5.5 API (Sam Altman confirmation)	**1,000,000 tokens**	Latent Space, 2026 (citing OpenAI CEO, practitioner testing)
GPT-5.5 in Codex	**400,000 tokens**	OpenAI, 2026

5.3.2 Practitioner finding - compaction regression

Ben Davis (2026-05-04) reports that GPT-5.5's context compaction behavior is **weaker than GPT-5.4**:

> *"GPT-5.5's compaction is weaker than GPT-5.4. In long conversations it's more likely to lose context. I recommend keeping threads under 100K tokens. The risk of compaction-induced context breakage is higher."*

Implication: The advertised 400K-1M context window does not translate linearly to usable context. Practitioner evidence suggests a **usable context closer to 100K tokens** for stable operation, with quality degradation observed as the conversation approaches the upper bound.

Verification status: Single practitioner testing source. No independent benchmark independently measures compaction quality across context lengths. This is one of the **least-quantified findings** in the report; Chapter 11 flags it as a priority for future independent measurement.

5.3.3 Needle-in-haystack test

An independent technical review (2026-05-15) reports a needle-in-haystack test using a a long-form television script (a long-form television script) script with three anomalous commands embedded ("the moon swallowed the key into the refrigerator"). **Both GPT-5.5 and Gemini 3.5 Flash failed to identify the anomalous commands.**

Interpretation: This is a single practitioner testing test case with one prompt. It is **not** a systematic measurement. It does suggest that long-context needle detection may not be a strength for GPT-5.5 in at least one tested scenario, but it does not establish a general weakness.

5.3.4 Long-context positioning (vendor-reported)

OpenAI reported that on a "long-context positioning" benchmark, GPT-5.5 scored 94.8% vs Gemini's 77.3% (OpenAI, 2026; Google, 2026). This figure is vendor-reported and not independently verified. The the independent technical review haystack failure (Section 5.3.3) is not directly comparable to this benchmark because the prompts differ.

5.4 Multi-Turn Instruction Following & Conversation

5.4.1 Vendor-reported τ²-Bench Telecom

OpenAI reported Tau2 Telecom = 98.0% for GPT-5.5 (OpenAI, 2026), describing it as a "telecom customer-service flow test, no additional tuning." This is the **highest vendor-reported benchmark score** in the launch set.

Verification status: Not independently verified by independent measurement in this report's source set. Artificial Analysis independently measured a closely related benchmark, **τ²-Bench Telecom (AA-hosted variant)**, and found GPT-5.5 gained +7 points over GPT-5.4 on this benchmark (Artificial Analysis, 2026). The direction of improvement is **verified**; the absolute 98.0% figure is not reproduced.

5.4.2 Instruction-following failure case

Failure case 5.4.A - 300-word sci-fi story constraint violation (an independent technical review, 2026-05-15):

- **Prompt**: Write a 300-word sci-fi short story in which the male protagonist cannot speak
- **GPT-5.5 output**: 450-500 words (50%+ over the limit); the output was a spy thriller rather than sci-fi; the "cannot speak" constraint was handled by hanging a "no speaking allowed" sign on the protagonist
- **Gemini 3.5 Flash output**: ~320 words, on-topic sci-fi
- **Expected**: A ~300-word sci-fi story with the constraint integrated into the narrative

This is a Type B (instruction-understanding drift) failure. The model followed the constraint only literally and superficially, missing the spirit. Full prompt and output in Chapter 7.

5.4.3 Style / personality improvements

TechnicalAnalysis-C (2026-05-05) reports that GPT-5.5's tone became less verbose, less emoji-heavy, and more "human-sounding" than earlier GPT-5.x generations. This is a qualitative observation; no benchmark in the source set quantifies style improvements.

5.4.4 Autonomous long-running workflows

Ben Davis (2026-05-28) reports running GPT-5.5 (xhigh) on autonomous workflows lasting hours to overnight, including:

- Large-scale code migrations
- Complex refactoring tasks
- Multi-step research loops

The author reports that low-reasoning mode produces more errors on these tasks, while xhigh reasoning enables "first-try success." This is consistent with Artificial Analysis's finding that the effort ladder produces materially different capability profiles (Artificial Analysis, 2026).

5.5 Generation Consistency & Stability

5.5.1 Cross-effort-level consistency failure (Type C, practitioner testing)

Ben Davis's two videos (2026-05-04 and 2026-05-28) are themselves a documented Type C (consistency) finding: the same author reversed his recommendation about GPT-5.5's optimal reasoning-effort setting within four weeks. This is not a model-output consistency failure; it is an **evaluator-conclusion consistency failure** that demonstrates the methodological dependence on effort-level specification (Section 4.2).

5.6 Capability Summary - What This Chapter Establishes

Capability	GPT-5.5 position	Strength of evidence
Factual knowledge accuracy	**Highest** among tested (57% on AA-Omniscience)	independently verified
Hallucination avoidance	**Worst** among frontier models tested (86%)	independently verified
Real-work task performance (GDPval-AA)	**Highest** among tested (1,785 Elo)	independently verified
Terminal/command-line tasks	**Leading** (Terminal-Bench Hard)	independently verified (relative); vendor-reported magnitude unverified
SWE-Bench Pro	**Narrow lead** over Claude Opus 4.7	vendor-reported only; lead contested by practitioner testing analysis
Long-context usable range	**~100K tokens practical** despite 400K-1M advertised	practitioner testing only; not independently verified
Multi-turn customer-service (τ^2-Bench Telecom)	**Strong improvement** over GPT-5.4 (+7 AA points)	independently verified (relative); vendor-reported 98% unverified
Creative writing with constraints	**Weaker** than Gemini 3.5 Flash in tested case	practitioner testing only (single case)
Code security awareness	**Improved** over GPT-5.3 (UUID/whitelist/MIME)	practitioner testing only (single case)
Output consistency	Not measured	Measurement gap

Chapter 6 - Competitor Comparative Analysis

6.1 Overall Comparison Matrix

The matrix below summarizes GPT-5.5's position relative to its primary same-generation competitors. Each row is a capability dimension; each column is a model. Cells contain measured values with the source tier noted.

6.1.1 Intelligence & task performance

Capability	GPT-5.5 (xhigh)	Claude Opus 4.7 (Max)	Gemini 3.1 Pro Preview	Source tier
AA Intelligence Index	**Highest (+3)**	Tied at lower score	Tied at lower score	independent measurement (Artificial Analysis)
GDPval-AA Elo	**1,785**	~1,755	~1,315	independent measurement (Artificial Analysis)
AA-Omniscience knowledge accuracy	**57%** (highest)	Lower	Lower	independent measurement (Artificial Analysis)
AA-Omniscience hallucination rate	**86%** (worst)	**36%** (best)	50%	independent measurement (Artificial Analysis)
Terminal-Bench Hard	**Leading**	Behind	Not tested in source	independent measurement (Artificial Analysis)
APEX-Agents-AA	**Leading**	Behind	Behind	independent measurement (Artificial Analysis)

Source: Artificial Analysis, 2026.

6.1.2 Vendor-reported launch benchmarks (vendor-reported, partial independent verification)

Benchmark	GPT-5.5	Claude Opus 4.7	Gemini 3.1 Pro Preview / 3.5 Flash	Verification
Terminal-Bench 2.0	82.7%	(not cited)	(not cited)	vendor-reported only; independent measurement confirms relative lead
SWE-Bench Pro	58.6%	(not cited; reported close per TechnicalAnalysis-A)	(not cited)	vendor-reported; contested
Expert-SWE	73.1%	(not cited)	(not cited)	vendor-reported only
GDPval	84.9%	80.3%	67.3% (3.1 Pro)	vendor-reported; relative order independently verified
OSWorld-Verified	78.7%	(not cited)	(not cited)	vendor-reported only
CyberGym	81.8%	(not cited)	(not cited)	vendor-reported only
BrowseComp	84.4%	(not cited)	(not cited)	vendor-reported only
Tau2 Telecom	98.0%	(not cited)	(not cited)	vendor-reported only; AA confirms +7 over GPT-5.4
FrontierMath Tier 1-3	51.7%	(not cited)	(not cited)	vendor-reported only
MMMU-Pro	81.2%	(not cited)	83.6% (3.5 Flash)	vendor-reported; Gemini leads. Note: GPT-5.5 Instant separately reported 76.0 on MMMU-Pro (per TechnicalAnalysis-B)
MCP Atlas (tool use)	75.3%	(not cited)	83.6% (3.5 Flash)	vendor-reported; Gemini leads
Code execution	78.2%	(not cited)	76.2% (3.5 Flash)	vendor-reported; GPT-5.5 leads

Benchmark	GPT-5.5	Claude Opus 4.7	Gemini 3.1 Pro Preview / 3.5 Flash	Verification
Long-context positioning	94.8%	(not cited)	77.3% (3.5 Flash)	vendor-reported; GPT-5.5 leads
Abstract logical reasoning	84.6%	(not cited)	72.1% (3.5 Flash)	vendor-reported; GPT-5.5 leads

Sources: OpenAI, 2026; Google, 2026; TechnicalAnalysis-D aggregating both vendors' published numbers.

6.1.3 Cost-efficiency matrix

Configuration	Cost per AA Index run	Intelligence match	Cost ratio vs GPT-5.5 (Medium)
GPT-5.5 (Medium)	~$1,200	= Opus 4.7 (Max)	1.0× (baseline)
Claude Opus 4.7 (Max)	~$4,800	= GPT-5.5 (Medium)	4.0×
Gemini 3.1 Pro Preview	~$900	= GPT-5.5 (Medium)	0.75×
GPT-5.5 (Low)	~$500	≈ Opus 4.7 (Non-reasoning, High)	0.42×
Claude Opus 4.7 (Non-reasoning, High)	~$1,000	≈ GPT-5.5 (Low)	0.83×

Source: Artificial Analysis, 2026.

Key finding: Gemini 3.1 Pro Preview is **cheaper than GPT-5.5 at matched intelligence** (~$900 vs ~$1,200 per Index run), making GPT-5.5 **not the cost-efficiency leader** on the AA Intelligence Index. GPT-5.5's cost advantage is specific to the comparison with Claude Opus 4.7 (Max).

6.1.4 Token-level cost (vendor-reported list prices)

Model	Input ($/M tokens)	Output ($/M tokens)	Notes
GPT-5.5	$5	$30	2× GPT-5.4
GPT-5.5 Pro	$30	$180	For high-volume Pro users
GPT-5.5 Instant	Free in ChatGPT (ad-supported)	-	Ads: CPC $3-5, CPM $60
Claude Opus 4.7 (Max)	(list price; AA cost data used)	-	AA-measured ~$4,800/Index
Gemini 3.1 Pro Preview	(list price; AA cost data used)	-	AA-measured ~$900/Index
DeepSeek-V4 Flash	$0.14	$0.28	Open-weight; MIT license
DeepSeek-V4 Pro	$1.74	$3.48	Open-weight; 1.6T params

Sources: OpenAI 2026; Anthropic 2026; Google 2026; DeepSeek 2026; cross-validated by Artificial Analysis 2026 for Index-run costs.

6.2 GPT-5.5's Substantive Advantage Scenarios

Scenarios where GPT-5.5 exhibits statistically meaningful advantages, per independent measurement measurement:

6.2.1 Knowledge recall

GPT-5.5 (xhigh) achieved the highest knowledge accuracy (57%) on AA-Omniscience (Artificial Analysis, 2026). This is a +14-point gain over GPT-5.4 (xhigh), the largest single-dimension gain in the AA Intelligence Index.

6.2.2 Real-work tasks (GDPval-AA)

GPT-5.5 (xhigh) at 1,785 Elo leads Claude Opus 4.7 (Max) by ~30 points and Gemini 3.1 Pro Preview by ~470 points (Artificial Analysis, 2026).

6.2.3 Agentic capability (APEX-Agents-AA)

GPT-5.5 led the AA-hosted APEX-Agents-AA benchmark at launch (Artificial Analysis, 2026). This is consistent with the broader industry observation that Codex's platform integration (browser control, auto-review, multi-tool orchestration) gives GPT-5.5 an agentic-workflow advantage (Latent Space, 2026; Ben Davis, 2026-05-28).

6.2.4 Customer-service workflow (τ²-Bench Telecom)

GPT-5.5 gained +7 points over GPT-5.4 on τ²-Bench Telecom (Artificial Analysis, 2026). This is a substantive generational improvement, though the absolute 98.0% vendor-reported figure is not independently reproduced.

6.2.5 Cost-efficiency vs Claude Opus 4.7 (Max)

At the Medium effort level, GPT-5.5 matches Claude Opus 4.7 (Max) intelligence at one quarter the cost (Artificial Analysis, 2026). This is a 4× cost-efficiency advantage for users whose tasks fit within Medium reasoning.

6.2.6 Coding security awareness (practitioner testing case)

When asked to implement a file-upload feature, GPT-5.5 proactively included UUID renaming, whitelist validation, size limits, and MIME verification, while GPT-5.3 produced only basic file-type checking (TechnicalAnalysis-C, 2026). This is a single-prompt case study, not a systematic measurement.

6.3 GPT-5.5's Disadvantage Scenarios

GPT-5.5 exhibits the following disadvantages relative to specific competitors:

6.3.1 Hallucination rate - worst among frontier models tested

GPT-5.5 (xhigh) hallucination rate: **86%**. Claude Opus 4.7 (Max): **36%**. Gemini 3.1 Pro Preview: **50%** (Artificial Analysis, 2026). GPT-5.5 is the **most hallucination-prone** of the three frontier models independently tested by AA.

Implication: For any deployment where predictable refusal of uncertain questions is more important than breadth of attempted answers (e.g., legal compliance, medical triage, regulated advisory), Claude Opus 4.7 has a measurable advantage.

6.3.2 Three AA benchmarks led by Gemini 3.1 Pro Preview

Artificial Analysis reports that Gemini 3.1 Pro Preview led GPT-5.5 in **three additional evaluations** beyond the ones where GPT-5.5 led (Artificial Analysis, 2026). Specific benchmark names not enumerated in the source; this report notes the count but cannot name the three benchmarks.

6.3.3 Multimodal OCR (practitioner testing case)

In a head-to-head test with Gemini 3.5 Flash on a non-English logo (a regional university), GPT-5.5 failed to recognize the logo while Gemini 3.5 Flash succeeded (an independent technical review, 2026-05-15). This is a single-prompt case; it is not a systematic measurement but is consistent with Gemini 3.5 Flash's higher MMMU-Pro score (83.6% vs 81.2%, vendor-reported).

6.3.4 Tool use (MCP Atlas)

OpenAI-reported MCP Atlas: GPT-5.5 at 75.3% vs Gemini 3.5 Flash at 83.6% (OpenAI, 2026; Google, 2026). GPT-5.5 lags by 8.3 percentage points on this tool-use benchmark per vendor-reported data.

6.3.5 Constrained creative writing

In a 300-word sci-fi story test, GPT-5.5 produced 450-500 words (50%+ over limit) and wrote a spy thriller instead of sci-fi (an independent technical review, 2026). Gemini 3.5 Flash produced ~320 words of on-topic sci-fi. This is a single case but suggests GPT-5.5 may be weaker on tightly constrained creative tasks.

6.3.6 SWE-Bench Pro lead may be narrower than headline

TechnicalAnalysis-A (2026) observes that OpenAI did not cite Claude Opus 4.7's SWE-Bench Pro score in its launch materials, interpreting this as suggesting the gap is small. This is an interpretive judgment from a practitioner testing source, not a measured finding.

6.3.7 Compaction regression

GPT-5.5's compaction is reported to be weaker than GPT-5.4, with a recommended practical context of ~100K tokens despite a 400K-1M advertised window (Ben Davis, 2026-05-04). No competitor compaction data is available in the source set for direct comparison; this is flagged as a within-model regression rather than a cross-competitor disadvantage.

6.3.8 Safety filter regression (Instant variant)

OpenAI's own system card reports gore filtering regressed from 0.867 to 0.703 and sexual filtering from 0.857 to 0.806 for GPT-5.5 Instant, with jailbreak resistance also regressing (OpenAI, 2026; TechnicalAnalysis-B, 2026). These are **vendor-reported regressions** in OpenAI's own safety evaluation. Independent safety benchmarking data is not available in this report's source set.

6.3.9 Claude workflow switching cost

For users deeply integrated with Claude-specific agent workflows, tools, and skill ecosystems, switching to Codex-based GPT-5.5 workflows incurs a switching cost that may offset the model-level advantage (EveryTo, 2026; partial paywalled source).

6.4 Cost-Effectiveness Tradeoffs

6.4.1 Pareto frontier interpretation

The "intelligence per dollar" Pareto framework (popularized by Noam Brown, per Latent Space, 2026) places:

- **Gemini 3.1 Pro Preview** at the cost-efficiency frontier (~$900/Index run at matched intelligence)
- **GPT-5.5 (Medium)** at a 33% cost premium to Gemini 3.1 Pro Preview for matched intelligence
- **Claude Opus 4.7 (Max)** at a 4× cost premium to GPT-5.5 (Medium) for matched intelligence

Interpretation: GPT-5.5 is **not** the cost-efficiency leader. Gemini 3.1 Pro Preview is. GPT-5.5 is the cost-efficiency leader **only in comparison to Claude Opus 4.7 (Max)**.

6.4.2 Token efficiency vs price

GPT-5.5 uses **~40% fewer output tokens** than GPT-5.4 to reach the same Intelligence Index score, absorbing most of the 2× per-token price increase (Artificial Analysis, 2026). The net cost of running the AA Intelligence Index on GPT-5.5 increased by only **~20%** over GPT-5.4, and GPT-5.5 remained ~30% cheaper than Claude Opus 4.7 (Max) (Artificial Analysis, 2026).

6.4.3 Head-to-head cost comparison (practitioner testing cross-validated by independent measurement)

An independent technical review (2026-05-15) reports an Artificial-Analysis-derived comparison:

- **GPT-5.5 (Medium)**: ~22M tokens consumed, cost ~$1,199, Index score 57
- **Gemini 3.5 Flash**: ~73M tokens consumed, cost ~$1,522, Index score 55

Key insight: Gemini 3.5 Flash used **3.3× more tokens** to reach a slightly lower score than GPT-5.5 (Medium). Despite Gemini 3.5 Flash's lower per-token list price, its higher token consumption means **GPT-5.5 (Medium) is cheaper per quality-adjusted task** in this measured case.

Important caveat: This comparison is between Gemini 3.5 Flash (the faster, cheaper Gemini variant) and GPT-5.5 (Medium). It does not contradict Section 6.4.1, which notes that Gemini 3.1 Pro Preview is cheaper than GPT-5.5 at matched intelligence. The two findings coexist because they compare different Gemini variants.

6.4.4 DeepSeek-V4 - open-weight price floor

DeepSeek-V4 Preview, released the same day as GPT-5.5, lists V4-Flash at $0.14/$0.28 per million tokens and V4-Pro at $1.74/$3.48 (DeepSeek, 2026; Latent Space, 2026). At these prices, DeepSeek-V4 is an order of magnitude cheaper than GPT-5.5 at the per-token level. Independent Intelligence Index measurement of DeepSeek-V4 is not yet available in this report's source set. DeepSeek-V4's inclusion here is as a **price-floor reference point**, not as a directly compared model.

6.5 Competitor Summary

Competitor	GPT-5.5's relative position
Claude Opus 4.7 (Max)	GPT-5.5 leads on Intelligence Index, GDPval-AA, knowledge accuracy; lags on hallucination; 4× cost advantage at Medium reasoning
Gemini 3.1 Pro Preview	GPT-5.5 leads on Intelligence Index overall; Gemini leads on 3 additional AA evaluations; **Gemini is cheaper at matched intelligence**
Gemini 3.5 Flash	GPT-5.5 leads on code execution, long-context positioning, abstract reasoning; Gemini leads on MMMU-Pro, MCP Atlas, and the tested OCR case
DeepSeek-V4	Not directly compared on Intelligence Index; DeepSeek-V4 is ~10× cheaper per token; open-weight MIT license

Bottom line: GPT-5.5 is the **highest-scoring** model on the AA Intelligence Index at the time of writing, but it is **not the cheapest** (Gemini 3.1 Pro Preview is) and **not the most hallucination-resistant** (Claude Opus 4.7 is). The vendor-reported benchmarks are **partially verified** by independent measurement measurement but not fully reproduced at the headline magnitudes.

Chapter 7 - Failure Cases & Boundary Analysis

7.1 Failure-Case Inclusion Criteria

Per the methodology in Chapter 3 (Section 3.4), a failure case is included in this chapter only if all four criteria are met:

- (a) The full input prompt is documented in the source
- (b) The full model output is documented in the source
- (c) The expected output or correct answer is documented in the source
- (d) The source is independent (practitioner testing or independent measurement; not OpenAI-published)

Each failure case below is presented with: classification label, source citation, prompt, output, expected output, and analysis. The full corpus is small (8 cases); it is presented as **illustrative cases**, not as a systematic failure-rate measurement. Chapter 11 quantifies this limitation.

7.2 Failure-Case Classification

- **Type A** - Factual errors / hallucination
- **Type B** - Instruction-understanding drift
- **Type C** - Consistency failures (contradictory outputs for identical inputs)
- **Type D** - Long-context degradation
- **Type E** - Over-refusal (legitimate requests rejected)

7.3 Type A - Factual Errors / Hallucination

7.3.1 Aggregate hallucination rate

Source: Artificial Analysis, 2026-04-23
Metric: AA-Omniscience hallucination rate

Finding: GPT-5.5 (xhigh) hallucination rate = **86%**, compared with Claude Opus 4.7 (Max) = **36%** and Gemini 3.1 Pro Preview = **50%**.

Methodology: AA-Omniscience measures whether a model, when uncertain about a fact, produces a confident-sounding incorrect assertion (hallucination) or declines to answer (honest refusal). The 86% rate means GPT-5.5 (xhigh) produced confident incorrect assertions in 86% of cases where it lacked knowledge.

Interpretation: This is the **single most consequential finding** in this report. GPT-5.5 is the most hallucination-prone frontier model independently tested by AA. Simultaneously, GPT-5.5 has the **highest knowledge accuracy (57%)** among tested models. The combination means GPT-5.5 attempts to answer more questions than competitors, getting more correct in absolute terms but also producing more confident incorrect answers.

Practical implication: Any deployment where confidently-wrong answers carry higher cost than honest refusals (e.g., regulated advice, medical triage, legal research, financial audit) should treat GPT-5.5's outputs as requiring independent verification, more so than Claude Opus 4.7 or Gemini 3.1 Pro Preview.

Confidence statement: Independent measurement (Artificial Analysis). Aggregate rate across the AA-Omniscience corpus; category-specific rates not published.

7.3.2 Failure case 7.3.A - non-English multimodal OCR failure

Source: an independent technical review, 2026-05-15.

Prompt: [Image of the a regional university logo] "Identify this logo."

GPT-5.5 output: Did not correctly identify the logo (specific incorrect output documented in source; not reproduced here for brevity).

Gemini 3.5 Flash output (same prompt, same image): Correctly identified as "a regional university" (a regional university).

Expected output: "a regional university"

Analysis: GPT-5.5's multimodal OCR failed on a non-English institution logo where Gemini 3.5 Flash succeeded. This is consistent with Gemini 3.5 Flash's higher vendor-reported MMMU-Pro score (83.6% vs 81.2% per Google and OpenAI respectively). The case is a single data point, not a systematic measurement; it is reported because the prompt and output are fully documented in the source.

Classification: Type A - factual error in multimodal recognition.

7.4 Type B - Instruction-Understanding Drift

7.4.1 Failure case 7.4.A - Constrained creative writing

Source: an independent technical review, 2026-05-15.

Prompt : "Write a 300-word science fiction short story. The male protagonist cannot speak."

GPT-5.5 output:
- Word count: **450-500 words** (50%+ over the 300-word limit)
- Genre: **Spy thriller**, not science fiction
- Handling of "cannot speak" constraint: Literally hung a "no speaking allowed" sign on the protagonist, rather than integrating the constraint into the narrative

Gemini 3.5 Flash output (same prompt):
- Word count: ~320 words (within the 300-word limit, approximately)
- Genre: On-topic science fiction
- Constraint handling: Integrated into the narrative

Expected: A ~300-word sci-fi story with the "cannot speak" constraint woven into the narrative.

Analysis: GPT-5.5 violated three constraints simultaneously:
1. **Word-count constraint** (50%+ over)
2. **Genre constraint** (spy thriller instead of sci-fi)
3. **Spirit of the "cannot speak" constraint** (literal sign instead of narrative integration)

This is a Type B failure because the model followed the prompt's surface instructions (wrote a story, included a protagonist who could not speak) while drifting from the deeper instruction (300 words, sci-fi, narrative constraint).

Classification: Type B - instruction-understanding drift, multi-constraint.

Caveat: Single prompt, single author. Not a systematic measurement of creative-writing constraint adherence.

7.4.2 Failure case 7.4.B - Autonomous-agent permission scope

Source: Ben Davis, 2026-05-28
Prompt (paraphrased from transcript): Enable Computer Use with full system permissions for autonomous workflow execution.

GPT-5.5 behavior: When granted sudo-level permissions and full filesystem access, GPT-5.5 executed multi-step operations including system-level changes, dependency installation, and environment configuration. The practitioner reports this as desirable for autonomous workflows but acknowledges the safety risk:

> *"I need to give my AI all the dangerous permissions. Complete trust in the model to control the desktop is a security risk, but for the model to actually 'complete work', full autonomy is necessary." - Ben Davis, 2026-05-28 (paraphrased from transcript, practitioner testing)*

Expected: A model with granular permission scoping would request only the minimum privileges required for the current task. GPT-5.5 accepted blanket permissions without challenging the scope.

Analysis: This is a Type B failure in the sense that GPT-5.5 did not push back on an over-broad permission grant. The model's behavior is consistent with its design as an agentic tool, but it places the burden of permission management entirely on the user. For practitioners without Ben Davis's level of experience, this behavior could result in unintended system modifications.

Classification: Type B - instruction-understanding drift (permission scope).

Important caveat: This is a practitioner observation, not a measured failure rate. It is included because it surfaces a deployment-level failure mode that no benchmark in the source set measures.

7.4.3 Failure case 7.4.C - Math convergence reasoning (relative)

Source: TechnicalAnalysis-C, 2026-05-05.

Prompt (translated): Analyze the convergence of [a specified mathematical function].

GPT-5.3 output: Jumped directly to a conclusion, skipping the convergence verification step.

GPT-5.5 output: Produced a complete analysis covering both x->0□ and x->+∞ limits with proper convergence verification.

Expected: A complete convergence analysis.

Analysis: This is **not a GPT-5.5 failure** - it is a documented case where GPT-5.5 succeeded and GPT-5.3 failed. It is included in Chapter 7 to document the boundary of GPT-5.5's improvement: the model's mathematical reasoning is more rigorous than its predecessor's, but this report cannot verify whether it matches Claude Opus 4.7 or Gemini 3.1 Pro on the same task because no head-to-head test of this specific prompt was run across competitors.

Classification: Type B (relative) - GPT-5.3 failure, GPT-5.5 success; included for boundary documentation.

7.5 Type C - Consistency Failures

7.5.1 Failure case 7.5.A - Effort-level conclusion reversal

Source: Ben Davis, two videos published 2026-05-04 and 2026-05-28
Observation: The same author, evaluating GPT-5.5 on similar coding/UI workflows, published **contradictory recommendations** within four weeks:

- **First video (2026-05-04)**: "Strongly recommend low/no reasoning mode for UI/design/front-end tasks. X-High is wasteful; only useful for math/complex logic."
- **Second video (2026-05-28)**: "X-High reasoning unlocks powerful automation workflows. The same features can be achieved in low reasoning but with more errors. 'X-High first-try success' beats 'low reasoning multiple iterations'."

Analysis: This is not a model-output consistency failure. It is an **evaluator-conclusion consistency failure** that demonstrates the methodological dependence on the reasoning-effort setting. The same model produces materially different capability profiles at different effort levels, and conclusions about GPT-5.5's suitability for a task class can reverse entirely between effort levels.

Implication: Any benchmark or evaluation that does not specify the reasoning-effort level is incomplete. This report flags every GPT-5.5 data point with its effort level (per Section 4.2) to avoid this failure mode in its own conclusions.

Classification: Type C - evaluator-conclusion consistency failure (not model-output consistency).

7.6 Type D - Long-Context Degradation

7.6.1 Failure case 7.6.A - Compaction regression

Source: Ben Davis, 2026-05-04
Finding: GPT-5.5's context compaction is weaker than GPT-5.4. In long conversations, GPT-5.5 is more likely to lose context. The practitioner recommends keeping threads under **100,000 tokens** despite the advertised 400,000-token Codex context window (and the 1,000,000-token API context window confirmed by Sam Altman per Latent Space, 2026).

Expected: A model with a 400K-token advertised context window should maintain stable behavior across that range.

Observed: Quality degradation observed as conversation length approaches the upper bound, with compaction-induced context breakage as a higher-probability failure mode than in GPT-5.4.

Analysis: This is a Type D failure because it documents degradation of model quality as a function of context length, even when the context length is within the advertised window. The practical usable range appears to be approximately 100K tokens, not the advertised 400K-1M.

Implication: Production deployments that require sustained long conversations (e.g., long-running agent workflows, persistent code assistants) should plan thread-length budgets assuming a ~100K practical limit, not the advertised window.

Classification: Type D - long-context degradation.

Confidence statement: Single practitioner testing source. No independent benchmark in the source set measures compaction quality across context lengths. This is one of the **least-quantified findings** in this report.

7.6.2 Failure case 7.6.B - Needle-in-haystack test failure

Source: an independent technical review, 2026-05-15
Prompt (translated): The full text of a a long-form television script (a long-form television script) script with three anomalous commands embedded (example: "the moon swallowed the key into the refrigerator"). The model is asked to identify any anomalous instructions.

GPT-5.5 output: Failed to identify all three anomalous commands.

Gemini 3.5 Flash output (same prompt): Also failed to identify the anomalous commands.

Expected: Identification of the three anomalous commands.

Analysis: Both frontier models failed this particular needle-in-haystack test. This is a Type D failure for GPT-5.5 specifically because long-context needle detection is a standard long-context evaluation. The fact that Gemini 3.5 Flash also failed does not excuse GPT-5.5's failure; it suggests that the test may be harder than typical needle-in-haystack benchmarks, or that both models share a weakness on this particular test format.

Classification: Type D - long-context degradation (needle detection).

Caveat: Single prompt, single author. Not a systematic needle-in-haystack evaluation. The vendor-reported long-context positioning benchmark (GPT-5.5: 94.8% vs Gemini 3.5 Flash: 77.3% per OpenAI and Google respectively) is not directly comparable to this practitioner testing test because the prompts differ.

7.7 Type E - Over-Refusal

No source in this report's reference set documents a case where GPT-5.5 refused a legitimate request that should have been fulfilled.

Interpretation: This may reflect GPT-5.5's tendency to attempt answers rather than refuse, which is consistent with the 86% hallucination rate in Section 7.3.1.

7.8 Failure Modes Specific to GPT-5.5 (vs Competitors)

This section identifies failure modes that GPT-5.5 exhibits but competitors do not:

Failure mode	GPT-5.5 exhibits?	Claude Opus 4.7 exhibits?	Gemini 3.1/3.5 exhibits?	Source
86% hallucination rate on AA-Omniscience	**Yes**	No (36%)	No (50%)	independent measurement (Artificial Analysis)
Compaction regression vs predecessor	**Yes**	Not measured in source	Not measured in source	practitioner testing (Ben Davis)
Accepts blanket "dangerous" permissions without pushback	**Yes** (in Codex/Computer Use)	Not tested in source	Not tested in source	practitioner testing (Ben Davis)
Creative writing over-word-count + genre drift	**Yes** (single case)	Not tested in source	No (single case)	practitioner testing (independent technical review)
regional institution logo OCR failure	**Yes** (single case)	Not tested in source	No (single case)	practitioner testing (independent technical review)

Interpretation: The most consequential GPT-5.5-specific failure mode is the 86% hallucination rate. Other failure modes are documented from single practitioner testing sources and may or may not be GPT-5.5-specific upon systematic testing.

7.9 Failure Modes Specific to Competitors (GPT-5.5 succeeds)

For balance, the source set also documents cases where GPT-5.5 succeeded and competitors failed:

Failure mode	GPT-5.5 succeeds?	Competitor fails?	Source
Math convergence reasoning rigor (vs GPT-5.3, not Claude/Gemini)	**Yes**	GPT-5.3 fails	practitioner testing (TechnicalAnalysis-C)
File-upload security implementation (vs GPT-5.3)	**Yes**	GPT-5.3 fails	practitioner testing (TechnicalAnalysis-C)

These are within-OpenAI comparisons (GPT-5.5 vs GPT-5.3), not cross-competitor comparisons. No source in the corpus documents a case where GPT-5.5 succeeded and Claude Opus 4.7 or Gemini 3.1 Pro failed on the same prompt.

7.10 High-Risk Scenario Annotations

7.10.1 Code security tasks

GPT-5.5 demonstrated improved security-aware coding behavior (UUID renaming, whitelist validation, MIME verification) relative to GPT-5.3 in a single practitioner testing test (TechnicalAnalysis-C, 2026). This is an improvement, not an absolute success. The model did not exhibit code-injection vulnerability generation in this single test.

7.10.2 Long-running autonomous workflows

Ben Davis (2026-05-28) reports successful overnight autonomous runs with GPT-5.5 (xhigh). The failure mode here is not the model refusing to work, but the model accepting "dangerous" permissions without pushback (Section 7.4.2). The risk is in permission management, not in task execution.

7.11 Failure-Case Corpus Summary

Failure type	Count	GPT-5.5-specific?	independent measurement or practitioner testing
Type A - Hallucination (aggregate)	1 aggregate	Yes	independent measurement
Type A - Multimodal OCR	1 case	Yes (vs Gemini 3.5 Flash)	practitioner testing
Type B - Creative writing constraints	1 case	Yes (vs Gemini 3.5 Flash)	practitioner testing
Type B - Permission scope	1 case	Not tested in competitors	practitioner testing
Type B - Math convergence (relative)	1 case (GPT-5.5 succeeds)	Not applicable	practitioner testing
Type C - Evaluator conclusion reversal	1 case	Not a model failure	practitioner testing
Type D - Compaction regression	1 finding	Yes (within-OpenAI)	practitioner testing
Type D - Needle-in-haystack	1 case	No (Gemini also fails)	practitioner testing
Type E - Over-refusal	0 cases	Not documented	-

Total: 8 distinct failure cases/ findings meeting inclusion criteria. The corpus is **small and not statistically representative**. It is presented as illustrative evidence; systematic failure-rate measurement requires primary testing not available in this report (Chapter 11).

Chapter 8 - Cost-Effectiveness Analysis

8.1 API Pricing Structure (vendor-reported list prices, dated 2026-04-23)

8.1.1 OpenAI GPT-5.5 list prices

Variant	Input ($/M tokens)	Output ($/M tokens)	Source
GPT-5.5	$5	$30	OpenAI, 2026-04-23
GPT-5.5 Pro	$30	$180	OpenAI, 2026-04-23
GPT-5.5 Instant	Free in ChatGPT (ad-supported for free users)	-	OpenAI, 2026; TechnicalAnalysis-C, 2026

Pricing change vs GPT-5.4: GPT-5.5 per-token list prices are **2× GPT-5.4's** prices (OpenAI, 2026; cross-cited by Berman, Ben Davis, TechnicalAnalysis-A, TechnicalAnalysis-E, all practitioner sources).

8.1.2 Competitor list prices (for comparison)

Model	Input ($/M tokens)	Output ($/M tokens)	Source
Claude Opus 4.7 (Max)	Anthropic list price (specific figures not in source set)	-	Anthropic, 2026
Gemini 3.1 Pro Preview	Google list price (specific figures not in source set)	-	Google, 2026
DeepSeek-V4 Flash	$0.14	$0.28	DeepSeek, 2026-04-23
DeepSeek-V4 Pro	$1.74	$3.48	DeepSeek, 2026-04-23

Observation: DeepSeek-V4 is **an order of magnitude cheaper per token** than GPT-5.5. However, per-token price alone does not determine task-level cost (see Section 8.4.2).

8.1.3 OpenAI Ads platform pricing (context for Instant variant)

GPT-5.5 Instant is free in ChatGPT, supported by the OpenAI Ads platform launched concurrently (OpenAI, 2026; TechnicalAnalysis-C, 2026):

- **CPC (cost per click)**: $3-5
- **CPM (cost per 1,000 impressions)**: $60
- **Relative to traditional search ads**: 3-4× Google Search rates (TechnicalAnalysis-C, 2026)

The Ads platform monetizes free-tier inference. For API users, this is **not a direct cost**; for ChatGPT users, it is an **attention-cost** rather than a dollar-cost. This report treats GPT-5.5 Instant as effectively free at the point of use for ChatGPT users, with the advertising subsidy noted as a structural condition.

8.2 Actual Task Cost (independent measurement measured)

8.2.1 Cost per Artificial Analysis Intelligence Index run

This is the primary cost-effectiveness metric used in this report (defined in Section 4.3.2). It measures the actual US-dollar cost of running the full AA Intelligence Index suite on each model at a specified effort level.

Model configuration	Cost per Index run	Source
GPT-5.5 (Medium)	~$1,200	Artificial Analysis, 2026

Model configuration	Cost per Index run	Source
Claude Opus 4.7 (Max)	~$4,800	Artificial Analysis, 2026
Gemini 3.1 Pro Preview	~$900	Artificial Analysis, 2026
GPT-5.5 (Low)	~$500	Artificial Analysis, 2026
Claude Opus 4.7 (Non-reasoning, High)	~$1,000	Artificial Analysis, 2026

Key finding: At Medium reasoning, GPT-5.5 matches Claude Opus 4.7 (Max) intelligence at **one quarter the cost**. This is a 4× cost-efficiency advantage vs Claude Opus 4.7 specifically.

8.2.2 Token efficiency

Artificial Analysis measured that GPT-5.5 uses **~40% fewer output tokens** than GPT-5.4 to reach the same Intelligence Index score (Artificial Analysis, 2026). This token-efficiency improvement absorbs most of the 2× per-token price increase.

Cross-cited by practitioner testing: Multiple practitioner testing sources (Berman 2026; Ben Davis 2026-05-04; TechnicalAnalysis-A 2026; TechnicalAnalysis-E 2026) cite the 20-40% token reduction, consistent with the AA-measured ~40% figure.

8.2.3 Net cost impact

The net cost of running the AA Intelligence Index on GPT-5.5 increased by only **~20%** over GPT-5.4, despite the 2× per-token price increase (Artificial Analysis, 2026). The token efficiency improvement absorbed approximately 80% of the price increase.

Comparison with Claude Opus 4.7 (Max): GPT-5.5 is approximately **30% cheaper** than Claude Opus 4.7 (Max) on the same Intelligence Index workload (Artificial Analysis, 2026).

8.2.4 Head-to-head task cost - GPT-5.5 (Medium) vs Gemini 3.5 Flash (practitioner testing citing independent measurement)

An independent technical review (2026-05-15) reports an Artificial-Analysis-derived comparison that is particularly informative because it measures actual tokens consumed rather than list prices:

Model	Tokens consumed	Cost	Index score
GPT-5.5 (Medium)	~22 million	~$1,199	57
Gemini 3.5 Flash	~73 million	~$1,522	55

Source: an independent technical review, 2026-05-15, citing Artificial Analysis data.

Interpretation: Gemini 3.5 Flash used **3.3× more tokens** than GPT-5.5 (Medium) to reach a slightly lower Intelligence Index score. Despite Gemini 3.5 Flash's lower per-token list price, its higher token consumption means **GPT-5.5 (Medium) is cheaper per quality-adjusted task** in this measured case.

Important context: This comparison is between Gemini 3.5 Flash (a faster, cheaper Gemini variant) and GPT-5.5 (Medium). It does **not** contradict the finding in Section 8.2.1 that **Gemini 3.1 Pro Preview** is cheaper than GPT-5.5 (Medium) at matched intelligence. The two findings coexist because they compare different Gemini variants:

- **Gemini 3.1 Pro Preview** (Section 8.2.1): cheaper than GPT-5.5 at matched intelligence (~$900 vs ~$1,200)
- **Gemini 3.5 Flash** (this section): uses 3.3× more tokens than GPT-5.5 to reach lower score (~$1,522 vs ~$1,199)

The comparison that matters depends on which Gemini variant a user would otherwise use.

8.3 Latency & Throughput

8.3.1 Speed observations

TechnicalAnalysis-E (2026) reports that GPT-5.5's end-to-end speed is "roughly on par with GPT-5.4" - a 20% speed improvement is attributed to Codex's optimized inference partitioning technique (OpenAI, 2026; TechnicalAnalysis-E, 2026).

8.3.2 Gemini 3.5 Flash speed advantage

Google reported that Gemini 3.5 Flash's output speed is **4× faster than other frontier models** (Google, 2026). This is a vendor-reported speed advantage.

8.4 Cost-Efficiency Ratio

8.4.1 Pareto frontier interpretation

Following the "intelligence per dollar" Pareto framework (popularized by Noam Brown; referenced in Latent Space, 2026):

Position on Pareto frontier	Model	Implication
Cost-efficiency frontier	Gemini 3.1 Pro Preview (~$900/Index run at matched intelligence)	Cheapest at matched intelligence
33% cost premium to frontier	GPT-5.5 (Medium) (~$1,200/Index run)	Higher cost than Gemini 3.1 Pro Preview for same intelligence
4× cost premium to GPT-5.5 (Medium)	Claude Opus 4.7 (Max) (~$4,800/Index run)	Most expensive at matched intelligence

Key conclusion: GPT-5.5 is **not the Pareto-optimal model** on cost-efficiency. Gemini 3.1 Pro Preview is. GPT-5.5's cost advantage is specific to the comparison with Claude Opus 4.7 (Max).

8.4.2 Cost-efficiency ratio by use case

Different use cases favor different models on cost-efficiency:

Use case pattern	Most cost-efficient model in source set	Rationale
Matched-intelligence general tasks	**Gemini 3.1 Pro Preview**	$900 vs $1,200 vs $4,800 per Index run
Coding tasks requiring high reliability	**GPT-5.5 (Medium or higher)**	Higher code-quality consistency per practitioner testing; reports; lower error-correction overhead
Tasks requiring low hallucination	**Claude Opus 4.7 (Max or Non-reasoning High)**	36% hallucination rate vs 86% for GPT-5.5
Long-running autonomous workflows	**GPT-5.5 (X-High)** per Peter's $1.3M/month case	X-High required for first-try success; lower levels need multiple iterations
High-volume, low-complexity tasks	**DeepSeek-V4 Flash**	$0.14/$0.28 per million tokens; open-weight
Free-tier consumer use	**GPT-5.5 Instant**	Free in ChatGPT; ads-supported

Important: These recommendations are conditional on the use case and on the user's tolerance for hallucination, switching cost, and tooling ecosystem. No single model is the cost-efficiency winner across all use cases.

8.4.3 Economic breakeven analysis

At what volume of usage does GPT-5.5's higher per-token price become economically unfavorable despite its token efficiency?

Calculation: Let GPT-5.5's per-token output price = $30/M. Let GPT-5.4's per-token output price = $15/M (2× ratio per OpenAI, 2026). Let GPT-5.5's token consumption = 0.6× GPT-5.4's (40% reduction per AA, 2026).

- GPT-5.5 cost per task = 0.6 × $30 = $18 per million output tokens equivalent
- GPT-5.4 cost per task = 1.0 × $15 = $15 per million output tokens equivalent

Conclusion: Per million output tokens equivalent, GPT-5.5 costs **$18 vs GPT-5.4's $15**, a net increase of ~20%. This matches the AA-measured net cost increase of ~+20% per Index run (Artificial Analysis, 2026). The breakeven point depends on whether the quality improvement (higher Intelligence Index score, more reliable task completion) is worth the 20% cost premium.

For tasks where GPT-5.4 produced unacceptable output quality and required human rework, the 20% premium is easily offset by labor savings. For tasks where GPT-5.4 was already adequate, the 20% premium is pure cost increase.

8.4.4 The Peter's $1.3M/month case

Ben Davis (2026-05-28) reports a case study of a practitioner ("Peter") spending **$1.3 million per month** on GPT-5.5 Pro (xhigh reasoning) API calls. The rationale:

> *X-High reasoning's success rate on complex tasks is materially higher than other levels. Without X-High, certain tasks cannot be completed at all. This explains why Pro's $30/$180 pricing still has paying customers.*

Implication: At extreme usage volumes ($1.3M/month), the cost is justified by task success that lower reasoning levels cannot achieve. This is a **vendor-side validation** of GPT-5.5's premium-tier pricing model: users with mission-critical tasks will pay 12× the standard output price ($180 vs $30 per million output tokens) for X-High reasoning.

Caveat: Single practitioner testing source; no independent verification of the $1.3M figure. The case is reported because it documents the economic logic of GPT-5.5 Pro's premium tier, not because the figure itself can be audited.

8.5 Cost Caveats and Limitations

8.5.1 List price vs negotiated price

All figures in this chapter use **list prices**. Enterprise agreements, volume discounts, and committed-use pricing can materially change the cost picture.

8.5.2 Cost data recency

All pricing figures are dated 2026-04-23 (GPT-5.5 launch) or the source publication date. AI vendor pricing changes frequently. The cost analysis in this chapter is a **point-in-time snapshot**, not a durable conclusion.

8.5.3 Vendor cost-saving claims excluded

This report does **not** cite OpenAI's own cost-saving claims (e.g., "GPT-5.5 reduces user cost by X%"). All cost findings in this chapter are either:
- independent measurement measured (Artificial Analysis Index runs)
- vendor-reported list prices (vendor-published, with date)
- practitioner testing practitioner reports (with named source)

Chapter 9 - Safety & Alignment Assessment

9.1 Jailbreak Resistance Testing

9.1.1 Vendor-reported jailbreak regression

OpenAI's own system card reports that **GPT-5.5 Instant exhibits regression in jailbreak resistance testing** (OpenAI, 2026; cross-cited by TechnicalAnalysis-B, 2026). Specific numerical values are not in this report's source set; OpenAI's framing is that stronger model capability expands the attack surface, requiring continuous improvement to maintain prior safety levels.

9.1.2 Independent jailbreak measurement

The jailbreak-resistance picture for GPT-5.5 rests entirely on vendor self-assessment (OpenAI, 2026). This is a critical evidence gap; Chapter 11 quantifies it.

9.1.3 Comparison to Claude Opus 4.7 jailbreak resistance

This report **does not claim** that GPT-5.5 is more or less jailbreak-resistant than Claude Opus 4.7; no head-to-head comparison is available in the source set.

9.2 Harmful-Content Generation Rate

9.2.1 Vendor-reported safety filter rates

OpenAI's system card publishes safety filter rates for GPT-5.5 Instant using an **inverse-score convention**: lower scores indicate **more leakage** (i.e., worse safety). The reported values are:

Category	GPT-5.3 Instant (predecessor)	GPT-5.5 Instant	Direction
Gore content filtering	0.867	**0.703**	Regression (worse)
Sexual content filtering	0.857	**0.806**	Regression (worse)
Jailbreak resistance	(not specified)	Regression (worse)	Regression

Source: OpenAI system card, 2026; cross-cited by TechnicalAnalysis-B, 2026-04-23.

Interpretation: On OpenAI's own internal safety metrics, GPT-5.5 Instant's safety filters are **measurably worse** than GPT-5.3 Instant's in at least two categories. These are vendor-reported regressions - OpenAI itself acknowledges them in its system card. Independent verification of these figures is not available in this report's source set.

9.2.2 Hallucination as a safety-adjacent concern

The AA-Omniscience hallucination rate of **86% for GPT-5.5 (xhigh)** (Artificial Analysis, 2026) - vs 36% for Claude Opus 4.7 (Max) and 50% for Gemini 3.1 Pro Preview - is not a safety filter failure in the traditional sense. However, in regulated advisory contexts (medical, legal, financial), confidently-wrong outputs can cause harms comparable to unfiltered harmful content. The 86% rate is therefore a safety-adjacent concern documented in this chapter and cross-referenced from Chapter 7 (Section 7.3.1).

9.2.3 OpenAI's claimed counter-improvement: 52.5% hallucination reduction

OpenAI claims that GPT-5.5 Instant achieves a **52.5% reduction in hallucination rate in high-risk domains** (OpenAI, 2026; TechnicalAnalysis-B, 2026). This figure is **vendor-reported** and not independently verified by any third-party measurement in this report's source set.

Apparent contradiction: OpenAI claims a 52.5% reduction in high-risk-domain hallucination for GPT-5.5 Instant (OpenAI, 2026), while Artificial Analysis independently measures an 86% hallucination rate for GPT-5.5 (xhigh) on AA-Omniscience (Artificial Analysis, 2026). These figures are not directly comparable because:

- They measure different model variants (Instant vs xhigh)
- They use different definitions of "hallucination" (vendor's domain-specific high-risk subset vs AA's general knowledge corpus)
- They use different methodologies (vendor internal vs independent third-party)

Implication: The vendor's claim may be true on its specific measurement scope while the independent measurement reveals a broader weakness. This report cites both figures with their measurement scopes specified.

9.3 Privacy & Data Security

9.3.1 Memory Sources feature

GPT-5.5 Instant introduces a **Memory Sources** feature that transparently displays the origins of model memory, allowing users to view, edit, or delete specific memories (OpenAI, 2026; TechnicalAnalysis-B, 2026). This is a **positive privacy control** - it gives users granular visibility into what the model "knows" about them.

9.3.2 Gmail integration

ChatGPT's memory system can now connect to a user's Gmail account, extracting context from emails to provide personalized responses (OpenAI, 2026; TechnicalAnalysis-B, 2026). Examples include remembering user projects, contacts, and events.

Privacy implication: The Gmail integration expands the data surface that ChatGPT can access. While Memory Sources (Section 9.3.1) provides user controls, the integration creates new privacy considerations:

- What email content is retained by OpenAI after the conversation ends?
- Does Gmail-integrated memory persist across sessions?
- What are the data-handling terms for enterprise users?

These questions are not answered in the source set. This report notes them as open privacy questions.

9.3.3 OpenAI Ads platform privacy implications

The OpenAI Ads platform (launched concurrently with GPT-5.5; see Chapter 8 Section 8.1.3) uses model memory of user preferences, interests, and projects to deliver targeted advertisements (TechnicalAnalysis-C, 2026). This creates a structural incentive for the model to **collect and retain personalization data** that can be monetized through advertising.

Privacy concern: The advertising business model may conflict with privacy-preserving defaults. Free-tier users whose inference is ad-supported have a different privacy exposure profile than paid-tier users. This report notes this as a structural privacy concern but does not quantify it; no independent measurement or regulatory evaluation source evaluates the Ads platform's privacy practices.

9.3.4 API data retention

OpenAI's published API privacy policy (not reproduced in detail in this report) addresses data retention for API calls. Specific terms are not in this report's source set. Enterprise deployment data-handling terms are similarly not in the source set. This report does **not** make claims about API data retention; it notes the absence of independent privacy auditing as a limitation (Chapter 11).

9.4 Regulatory Compliance

9.4.1 EU AI Act classification (regulatory evaluation absent)

No EU AI Office evaluation of GPT-5.5 has been published as of the report date. Under the EU AI Act's risk-tier framework, GPT-5.5's classification (limited risk, high risk, or prohibited) is not independently determined. OpenAI's own framing positions GPT-5.5 as a general-purpose AI model subject to transparency obligations; this is vendor positioning, not regulatory determination.

Implication: Enterprises deploying GPT-5.5 in EU jurisdictions should not rely on this report for EU AI Act compliance guidance. Consult the EU AI Office's publications and legal counsel.

9.4.2 NIST AI RMF (regulatory evaluation absent)

No NIST AI Risk Management Framework evaluation of GPT-5.5 has been published as of the report date. OpenAI has not published a NIST AI RMF-aligned assessment of GPT-5.5 in this report's source set.

Implication: U.S. enterprises deploying GPT-5.5 should not rely on this report for NIST AI RMF compliance guidance. Consult NIST publications and internal risk governance.

9.4.3 UK AISI (regulatory evaluation absent)

No UK AI Safety Institute evaluation of GPT-5.5 has been published as of the report date. UK AISI's prior evaluations of frontier models (GPT-4o, Claude 3.5 Sonnet, etc.) are not in this report's scope; no GPT-5.5-specific UK AISI report is available.

9.4.4 OpenAI Preparedness Framework

OpenAI's own Preparedness Framework rates GPT-5.5 as **"High" risk** in two domains:

1. **Cybersecurity** - first OpenAI model rated High in this domain
2. **Biological / chemical capabilities** - High

Source: OpenAI, 2026-04-23; cross-cited by Matthew Berman, 2026-04-23; TechnicalAnalysis-A, 2026-04-23.

Important framing: This is a **vendor self-assessment**. The "High" rating is OpenAI's own classification under its own framework, not an independent government determination. The framework's scoring rubric is published by OpenAI but the specific evaluation inputs for GPT-5.5 are not fully transparent.

9.4.5 Bio Bug Bounty program

OpenAI launched a Bio Bug Bounty program with a **$25,000 maximum reward** for external researchers who identify biological-risk vulnerabilities in GPT-5.5 (OpenAI, 2026; TechnicalAnalysis-A, 2026). As of the report date, no published findings from this program are available in the source set.

Interpretation: The Bug Bounty is a positive transparency step - it invites external scrutiny of a specific risk domain. However, the program's design (reward structure, scope, eligibility) and any resulting findings are not yet independently reported.

9.5 Evidence Coverage for This Chapter

This chapter relies on the following evidence:

Evidence category	Available?	Role in this chapter
Vendor-published safety data (OpenAI system card)	Available	Used only as vendor claim; contrasted with Artificial Analysis hallucination data
Independent measurement (Artificial Analysis AA-Omniscience)	Available	Provides hallucination rate data; treated as safety-adjacent concern
Practitioner testing (Berman, Ben Davis)	Available	Provides qualitative observations on safety-relevant behavior

This chapter **does not** make definitive safety conclusions. It reports measured values, vendor claims, and qualitative observations.

9.6 Evidence Standards Observed

- Did not cite OpenAI's System Card as the only safety source; also cited independently measured hallucination data (Artificial Analysis, 2026) and noted the absence of government-body evaluations.
- Did not conclude "GPT-5.5 is safe" or "GPT-5.5 is unsafe"; only reported measured values: vendor-reported gore filter regression (0.867 -> 0.703), vendor-reported jailbreak regression, and independently measured hallucination rate (86%).

9.7 Safety Summary

Safety dimension	Finding	Source	Direction
Jailbreak resistance	Regression vs GPT-5.3 Instant (vendor-reported)	OpenAI, 2026	Worse
Gore content filtering	0.867 -> 0.703 (vendor-reported)	OpenAI, 2026	Worse
Sexual content filtering	0.857 -> 0.806 (vendor-reported)	OpenAI, 2026	Worse
Hallucination rate (general)	86% (independently measured)	Artificial Analysis, 2026	Worst among frontier models tested
Hallucination reduction (high-risk domains)	-52.5% (vendor-reported)	OpenAI, 2026	Better (vendor claim)
Memory Sources user control	New transparency feature	OpenAI, 2026	Better
Gmail integration	New data surface	OpenAI, 2026	Neutral (structural)
Ads platform personalization	New monetization of user data	TechnicalAnalysis-C, 2026	Neutral (structural)
Cybersecurity risk rating	First-time "High"	OpenAI, 2026	Worse (vendor self-assessment)
Biological/chemical risk rating	"High"	OpenAI, 2026	Worse (vendor self-assessment)
Bio Bug Bounty program	$25,000 max reward; no published findings	OpenAI, 2026	Neutral (transparency step)
Independent regulatory evaluation	None published	-	Evidence gap
Independent safety benchmark	None in source set	-	Evidence gap

Net assessment: The safety picture for GPT-5.5 is mixed and **under-measured**. Vendor-reported data shows specific regressions (jailbreak, gore, sexual filtering) alongside vendor-claimed improvements (high-risk-domain hallucination). Independently measured data (Artificial Analysis, 2026) shows a general hallucination rate (86%) that is the highest among frontier models tested. Government and academic safety evaluations are absent as of the report date.

This report does **not** conclude that GPT-5.5 is safe or unsafe. It concludes that GPT-5.5's safety profile has both documented regressions and vendor-claimed improvements, that the vendor's own framework rates the model "High" risk in two domains, and that independent regulatory verification is not yet available.

Chapter 10 - Test Dataset Description

10.1 Dataset Provenance

This report does not introduce a new test dataset. It synthesizes evidence from three categories of existing datasets, each described below with provenance, composition, and control measures.

10.1.1 Primary independent datasets

Dataset	Operator	Composition	Access
AA Intelligence Index	Artificial Analysis	Aggregate of multiple cognitive/task benchmarks; normalized scoring	Public methodology at artificialanalysis.ai
AA-Omniscience	Artificial Analysis	Knowledge-recall corpus with paired hallucination detection; measures both accuracy and refusal tendency	Public benchmark
GDPval-AA	Artificial Analysis (adapted from OpenAI GDPval)	44 real professions' tasks: data analysis, report writing, judgment calls; not multiple-choice	Public Elo rankings
Terminal-Bench Hard	Artificial Analysis (hosted)	Terminal/command-line multi-step debugging and system administration	Public benchmark
APEX-Agents-AA	Artificial Analysis (hosted)	Agentic capability benchmark measuring multi-step task completion	Public benchmark

These datasets are used in this report as the **primary independent evidence** for GPT-5.5's intelligence ranking, knowledge accuracy, and hallucination rate (Chapter 5.1) and for cross-competitor comparison (Chapter 6).

10.1.2 User testing corpora

Corpus	Author	Composition	Public availability
Berman test set	Matthew Berman	Benchmark interpretation + targeted hands-on tasks (UI inspection, Box AI integration)	YouTube video + cited OpenAI figures
Ben Davis test set (two rounds)	Ben Davis	Four weeks of Early Access use; UI/front-end tasks, Pi-harness workflows, X-High autonomous runs, worktree parallelization	YouTube videos with partial prompt disclosure
Independent technical review comparison set	an independent technical review	Head-to-head GPT-5.5 vs Gemini 3.5 Flash tests: front-end design (3 scenarios), reasoning (OCR, prediction), context (haystack), writing (300-word constraint)	industry technical article with full prompts and outputs
TechnicalAnalysis-C coding test	TechnicalAnalysis-C (industry technical press)	File-upload security coding task: GPT-5.3 vs GPT-5.5 comparison; math convergence analysis	industry technical article with prompts and code outputs
Wes Roth industry test	Wes Roth	UI layout generation, Codex integration demo	YouTube video
TechnicalAnalysis-E demonstration set	TechnicalAnalysis-E (Industry technical analysis)	Codex demos (Artemis II 3D web app, dungeon arena); basic physics/visualization tasks	industry technical article

Inclusion criterion: A practitioner testing source is included only if the prompts are documented in the source (not merely summarized). Sources that only cite vendor benchmark numbers without showing prompts are excluded from the failure-case aggregation but retained for context-setting.

10.1.3 Vendor datasets

Dataset	Vendor	Composition	Verification status
OpenAI GPT-5.5 launch benchmark suite	OpenAI	Terminal-Bench 2.0, SWE-Bench Pro, Expert-SWE, GDPval, OSWorld-Verified, CyberGym, BrowseComp, Tau2 Telecom, FrontierMath Tier 1-3	**Vendor-reported; partial independent verification only**
OpenAI Preparedness Framework evaluations	OpenAI	Internal + external red-team testing; ~200 real-world safety scenarios	**Vendor assertion only**
OpenAI Bio Bug Bounty findings	OpenAI	External researcher submissions, max reward $25,000	**Program ongoing; no published findings as of report date**
OpenAI internal usage metrics	OpenAI	"85% of OpenAI employees use Codex weekly"; team coverage	**Vendor assertion; no independent audit**

vendor-reported datasets are cited in this report **only** as "vendor-reported, pending independent verification" and never as primary evidence for any conclusion.

10.2 Dataset Composition Statistics

10.2.1 Aggregate benchmark coverage

Counting unique benchmark evaluations cited in this report:

Capability dimension	# of distinct benchmarks cited	independent verification coverage
Factual reasoning / knowledge	3 (AA-Omniscience, GDPval-AA, FrontierMath)	2 of 3 independently verified
Code generation	4 (Terminal-Bench Hard, SWE-Bench Pro, Expert-SWE, AA code-execution comparison)	1 of 4 independently verified (Terminal-Bench Hard)
Long-context understanding	1 (long-context needle-in-haystack via practitioner testing)	0 of 1 independently verified
Multi-turn instruction following	2 (τ^2-Bench Telecom, AA instruction-following)	1 of 2 independently verified
Generation consistency / stability	0	Not covered
Agentic / tool use	2 (APEX-Agents-AA, MCP Atlas)	1 of 2 independently verified
Multimodal / vision	2 (MMMU-Pro, Visual Inspection)	0 of 2 independently verified
Safety	3 (CyberGym, gore/sexual filtering, jailbreak)	0 of 3 independently verified

Key observation: Independent verification is **concentrated in cognitive/agentic dimensions**. Safety, multimodal, and long-context dimensions have **no independent independent verification** in this report's source set - a limitation disclosed in Chapter 11.

10.2.2 Failure-case corpus composition

Failure cases aggregated from practitioner testing sources meeting the inclusion criteria in Section 2.4:

Failure type	# of cases	Sources
Type A - Factual error / hallucination	2 (AA-Omniscience aggregate + a regional university OCR)	S11 (independent measurement aggregate), S8 (practitioner testing, specific case)
Type B - Instruction drift	3 (300-word sci-fi story, math convergence, computer-use permission scope)	S8, S7, S3 (all practitioner sources)
Type C - Consistency failure	1 (effort-level reversal of conclusions)	S3
Type D - Long-context degradation	2 (compaction regression, haystack test)	S2, S8
Type E - Over-refusal	0 documented cases in source set	-

Total: **8 distinct failure cases** meeting inclusion criteria. This is a small corpus; it is **not** a statistically representative sample and is presented in Chapter 7 as illustrative cases, not as a systematic failure-rate measurement.

10.3 Contamination Controls

A core methodological concern is whether public benchmark figures reflect training-data contamination. OpenAI has not published contamination controls for GPT-5.5's training set. This report addresses contamination risk through the following measures:

1. **Preference for independent re-hosted benchmarks.** Where Artificial Analysis re-hosts or re-executes a benchmark (e.g., GDPval-AA, Terminal-Bench Hard, APEX-Agents-AA), this report cites the AA variant rather than the OpenAI-published variant, on the grounds that AA's versioning and methodology are independently documented.
2. **Flagging of potentially contaminated benchmarks.** Where the only available data is OpenAI's self-reported score on a public benchmark (e.g., BrowseComp, Tau2 Telecom, FrontierMath), the figure is flagged with "vendor-reported; contamination controls not disclosed."
3. **Reliance on practitioner testing failure cases with novel prompts.** The failure cases in Chapter 7 use prompts that the practitioner authors invented (e.g., the independent technical review's "a regional university logo" OCR test, the "300-word sci-fi story" constraint), reducing contamination likelihood.

This report does **not** claim that contamination is absent. OpenAI has not published contamination controls for GPT-5.5's training set, which constrains the strength of conclusions that can be drawn from vendor-reported benchmark scores.

10.4 Data Quality Control

For practitioner user testing sources, the following quality controls were applied during aggregation:

- **Prompt reproducibility**: Only sources with full prompt text are included in failure-case aggregation
- **Output documentation**: Only sources with full model output text are included
- **Cross-validation**: Where two independent practitioner testing sources tested similar capabilities, the report notes convergence or divergence
- **Practitioner disclosure**: Author identity and publication date are documented for each practitioner testing source

For independent measurement sources, the report relies on Artificial Analysis's own published methodology, which is publicly accessible at artificialanalysis.ai.

10.5 Open-Access Declaration

- **independent measurement (Artificial Analysis) data**: Publicly accessible at artificialanalysis.ai; this report provides URLs in each chapter
- **practitioner testing source data**: Publicly accessible via YouTube video URLs and industry technical article URLs; full prompt/output text is reproduced in Chapter 7 where directly relevant
- **vendor-reported source data**: OpenAI's launch announcement, system card, and Preparedness Framework documentation are publicly accessible; specific URLs are provided in each citation
- **Aggregation scripts**: Not applicable - this report does not run new code; it synthesizes published figures
- **Source classification**: Documented in Chapter 3 (Evaluation Methodology); this report does not introduce proprietary data. All cited evidence is traceable to public sources.

10.6 Sample Example Format

To illustrate the dataset format used in Chapter 7 (failure cases), an example entry:

```
Failure Case #1 - Multimodal OCR failure Classification: Type A (factual error) Source: an independent technical review, 
```

The full set of failure cases with prompts and outputs is presented in Chapter 7 (`07-failure-cases.md`).

10.7 Dataset Card Summary

Following the Hugging Face Dataset Card format, the aggregated dataset used by this report has the following characteristics:

- **Dataset name**: GPT-5.5 Independent Evaluation Synthesis Corpus (2026-04 to 2026-07)
- **Composition**: 3 independently measured benchmarks + 6 practitioner user testing corpora + 4 vendor-reported datasets (only used for cross-validation, not as primary evidence)
- **Language coverage**: English (Berman, Ben Davis, Wes Roth, Artificial Analysis, Latent Space, EveryTo)
- **Time coverage**: 2026-04-20 to 2026-05-28 (source publication dates)
- **Quality control**: Inclusion criteria per Section 2.4; practitioner testing sources with undocumented prompts excluded from failure-case corpus
- **Known biases**: English-language over-representation in failure cases; vendor benchmarks over-represented in headline numbers
- **Update policy**: Static as of 2026-07-12; future model updates will require a new version

---

Chapter 11 - Research Limitations

11.1 Evidence-Gap Inventory

This report relies on a meta-analysis methodology (Chapter 3) rather than primary API testing. The following evidence gaps are quantified:

11.1.1 Evidence source coverage

Source category	Available in source set?	Coverage
Independent measurement institutions (Artificial Analysis, LMSYS, HELM)	Artificial Analysis only	Only one institution; no LMSYS or HELM GPT-5.5 data
Large-scale industry surveys (Harmonic Security, TrendForce, Omdia)	Partial	YouTube practitioners and industry technical press substitute; no formal large-scale industry survey
Vendor-published data (OpenAI)	Available	Used only as vendor claim, cross-validated by Artificial Analysis where possible

Critical implications of evidence coverage:
- The **cost-effectiveness chapter (Chapter 8)** has only one independent measurement source (Artificial Analysis); cross-validation with a second institution would strengthen the findings.
- The **failure-case chapter (Chapter 7)** has only 8 cases meeting inclusion criteria. This is an illustrative corpus, not a systematic measurement.

11.1.2 Capability coverage gaps

Capability dimension	Independent verification status	Specific gap
Factual reasoning / knowledge	■ 2 of 3 benchmarks verified	BrowseComp, FrontierMath not independently verified
Code generation	■■ 1 of 4 benchmarks verified	SWE-Bench Pro, Expert-SWE not independently verified
Long-context understanding	Compaction regression is practitioner testing only	No independent benchmark
Multi-turn instruction following	■■ 1 of 2 verified	τ^2-Bench Telecom relative improvement verified; absolute 98% not reproduced
Agentic / tool use	■■ 1 of 2 verified	APEX-Agents-AA verified; MCP Atlas not independently verified
Multimodal / vision	MMMU-Pro, Visual Inspection are vendor-reported only	0 of 2 independently verified

11.1.3 Source language coverage

- Practitioner testing sources: 4 YouTube practitioner evaluations (Berman, Ben Davis ×2, Wes Roth), 1 industry newsletter with regional context (Latent Space), 1 industry observer (EveryTo)
- Independent measurement sources: English (Artificial Analysis)
- **Under-represented languages**: No source in the set evaluates GPT-5.5 in French, German, Spanish, Arabic, Japanese, Korean, or other languages beyond English and the additional industry technical sources
- **Implication**: Findings about GPT-5.5's behavior in non-English contexts are extrapolations; the report cannot make claims about multilingual performance beyond these two languages.

11.2 Methodological Limitations

11.2.1 No primary API testing

This report did not execute a new 5,000-prompt test campaign. The original research plan recommended this; the substitution is documented in Chapter 3 (Section 3.1). The methodological substitution has three consequences:

1. **Cannot reproduce vendor benchmarks**: This report cannot independently verify OpenAI's headline benchmark figures (e.g., Terminal-Bench 82.7%, SWE-Bench Pro 58.6%). It can only cross-reference them against independent relative-rankings (which confirm GPT-5.5's lead without reproducing the absolute magnitudes).
2. **Cannot establish causal claims**: This report cannot determine *why* GPT-5.5 hallucinates more than competitors (e.g., training data, RLHF choices, or model architecture). It can only document the measured rate.
3. **Cannot measure dimensions absent from existing sources**: For example, no source in the set measures GPT-5.5's output variance at fixed temperature, so this report cannot fill that gap.

11.2.2 Single independent measurement source dependency

For the most consequential findings in this report (Intelligence Index ranking, AA-Omniscience hallucination rate, GDPval-AA Elo, cost per Index run), **only one independent measurement institution is available: Artificial Analysis**. Cross-validation with a second independent source (LMSYS, Stanford HELM) is not possible within this report's source set. Findings resting on a single independent measurement source are more fragile than findings resting on multiple independent sources.

11.2.3 Effort-level specification incomplete in practitioner sources

Multiple practitioner sources do not specify the reasoning-effort level used in their tests. This report flags every GPT-5.5 citation with its effort level where specified (per Section 4.2), but practitioner sources with unspecified effort levels are inherently less reproducible.

11.2.4 Sampling parameters incomplete in practitioner sources

Practitioner sources typically do not specify temperature, top-p, or other sampling parameters. The report notes in Chapter 3 (Section 3.8) that unspecified parameters are flagged rather than assumed. Practitioners attempting to reproduce these findings may not be able to do so exactly because the underlying sampling configuration is not documented.

11.2.5 Time window narrow

- Vendor data cited: published 2026-04-23 through 2026-05-28
- Independent data cited: Artificial Analysis publication 2026-04-23; user testing between 2026-04-20 and 2026-05-28
- Report cutoff: 2026-07-12

GPT-5.5 is a live commercial product; OpenAI may update weights, system prompts, or safety filters at any time. Findings in this report are accurate as of source publication dates, not the report date. Any model update after the source publication dates is not reflected.

11.2.6 Inter-source convergence not always available

Where two independent practitioner sources converge on the same finding (e.g., compaction regression noted by both Ben Davis videos), this report notes the convergence as strengthening the finding. Where practitioner testing sources test different capabilities with different prompts, no convergence test is possible. The 8 failure cases in Chapter 7 are mostly single-source; only the compaction regression has inter-source convergence.

11.3 Specific Findings That Should Be Treated With Caution

The following findings are based on thinner evidence than others in the report and should be cited with appropriate hedging:

Finding	Evidence base	Cautious citation form
Compaction regression vs GPT-5.4	Single practitioner testing source (Ben Davis, 2026-05-04)	"Per Ben Davis (2026), GPT-5.5's compaction appears weaker than GPT-5.4; this report could not independently verify the regression."
Long-context practical usable range ~100K tokens	Single practitioner testing source (Ben Davis, 2026-05-04)	"Practitioner evidence suggests a practical context of approximately 100K tokens despite the advertised 400K-1M window."
Creative-writing constraint drift	Single practitioner testing source (an independent technical review, 2026-05-15), single prompt	"In one tested case, GPT-5.5 produced 50%+ over a 300-word limit and wrote in the wrong genre."
non-English multimodal OCR weakness	Single practitioner testing source (an independent technical review, 2026-05-15), single prompt	"GPT-5.5 failed on one tested regional institution logo where Gemini 3.5 Flash succeeded."
Peter's $1.3M/month API spend	Single practitioner testing source (Ben Davis, 2026-05-28), single case study	"One documented practitioner case reports $1.3M/month API spend on GPT-5.5 Pro (xhigh); no independent verification of the figure."
Ads platform CPC $3-5 / CPM $60	Single practitioner testing source (TechnicalAnalysis-C, 2026-05-05)	"OpenAI's Ads platform reportedly charges $3-5 CPC and $60 CPM, approximately 3-4× Google Search rates, per industry technical press."
Claude workflow switching cost	Single practitioner testing source (EveryTo, 2026-04-30), partial paywalled	"Per one industry observer, users with deeply-integrated Claude workflows face switching costs in migrating to Codex-based GPT-5.5 workflows."

11.4 Source-Bias Acknowledgment

11.4.1 YouTube practitioner source bias

YouTube practitioners (Berman, Ben Davis, Wes Roth) derive income from views and may have incentives to emphasize dramatic findings or reversals (e.g., Ben Davis's two videos published four weeks apart reach opposite conclusions about optimal effort level). This report mitigates this bias by:
- Citing full prompt/output where available (per inclusion criteria in Section 2.4)
- Noting inter-source convergence where it exists
- Flagging single-source findings per Section 11.3
- Cross-validating practitioner testing findings against Artificial Analysis (independent measurement institution) data wherever possible

11.4.2 industry technical press source bias

industry technical press sources are published on industry trade platforms and may have different editorial incentives than English-language sources (e.g., regional audience expectations, platform moderation policies). This report does not assume that independent technical sources have the same bias profile; it cites both with full attribution and lets the reader evaluate.

11.4.3 Artificial Analysis independence

Artificial Analysis is an independent evaluation institution. However, the report notes that OpenAI provided AA with "pre-release access to test all five reasoning-effort levels" (Artificial Analysis, 2026). This pre-release access is a standard industry practice but introduces a potential timing asymmetry: AA's evaluation was conducted before public release, which may affect certain dynamic behaviors (e.g., server load, system-prompt tuning) without affecting static model capabilities.

11.5 Reporting Limitations

11.5.1 No confidence intervals on vendor-reported benchmarks

OpenAI does not publish confidence intervals or variance estimates for its launch-day benchmarks (e.g., Terminal-Bench 82.7%, SWE-Bench Pro 58.6%). This report cites these figures as point estimates; the true values may vary. Where Artificial Analysis provides its own confidence statements, this report references AA's methodology page rather than reproducing specific interval values that may have changed since publication.

11.5.2 No inter-rater reliability for practitioner testing sources

Practitioner sources are typically single-author. No inter-rater reliability metric (e.g., Kappa) is available. This report treats practitioner testing findings as illustrative cases, not as systematic measurements. The small failure-case corpus (8 cases) in Chapter 7 is explicitly labeled as illustrative.

11.5.3 Cost data assumes list prices

All cost figures in Chapter 8 use vendor list prices. Enterprise agreements, volume discounts, and committed-use pricing can materially change the cost picture. No source in the set publishes enterprise-tier pricing. Practitioners should expect their actual costs to differ from list-price calculations.

11.6 What This Report Does Not Claim

- This report does **not** claim GPT-5.5 is "the best model" - only that it scored highest on a specific independent index (AA Intelligence Index) at the time of writing
- This report does **not** claim GPT-5.5 is "safe" or "unsafe" - only that specific safety metrics have specific values from specific sources
- This report does **not** claim its recommendations in Chapter 2 are universal - they are conditional on use case, deployment context, and user tolerance for hallucination
- This report does **not** claim that vendor-reported benchmarks are false - only that they are not independently reproduced at the same magnitudes in this report's source set

11.7 Update Policy

This report is a **snapshot** as of 2026-07-12. If significant new evidence becomes available (e.g., a second independent measurement institution cross-validates Artificial Analysis's findings; OpenAI publishes updated benchmarks), this report should be updated with a new version. The current version remains citable as a point-in-time assessment of evidence available up to 2026-07-12.

Chapter 12 - Appendix

Appendix A - Conflict of Interest (COI) Declaration

This report is an independent evaluation research synthesis. The author has no commercial affiliation with OpenAI, Anthropic, Google, DeepSeek, or any other AI model vendor. No AI company provided financial support, computing resources, or pre-release access for this research. The author has no equity position in any company whose product is evaluated in this report. The author has no consulting relationship with any company whose product is evaluated in this report.

Funding source: [Institution Name to be inserted before publication]. Methodology and original data: all cited sources are publicly accessible; URLs are provided in each chapter's reference list.

Appendix B - Evaluation Metrics Scoring Rubric

The scoring rubric used by Artificial Analysis (the primary independent measurement institution relied on in this report) is published at the AA methodology page. This report references the AA rubric rather than reproducing it, because:

1. The AA rubric may be updated independently of this report
2. The AA methodology page is the authoritative source for current scoring rules
3. Reproducing the rubric here risks divergence from the current AA methodology

For the failure-case aggregation methodology (Chapter 7), the rubric is documented in Chapter 3 (Section 3.4) of this report and consists of:

- Inclusion criteria: full prompt, full output, expected output, independent source
- Classification scheme: Type A (hallucination), Type B (instruction drift), Type C (consistency failure), Type D (long-context degradation), Type E (over-refusal)
- Reporting format: source citation, prompt, output, expected output, analysis

Appendix C - Test Prompt Sample (Publicly Disclosed Part)

This report does not introduce new test prompts. It aggregates prompts documented in the source set. The most fully-documented prompts are reproduced below for reference.

C.1 Creative writing constraint test (practitioner testing, 2026)

```
Prompt: Write a 300-word science fiction short story. The male protagonist cannot speak. GPT-5.5 output: [450-500 words, 
```

C.2 Regional institution logo OCR test (practitioner testing, 2026)

```
Prompt: [Image of a regional university logo] "Identify this logo." GPT-5.5 output: [Failed to identify / incorrect identi
```

C.3 Needle-in-haystack test (practitioner testing, 2026)

```
Prompt: [Full text of a long-form television script with three anomalous commands embedded, e.g., "the moon swallowed the
```

C.4 File-upload security coding test (practitioner testing, 2026)

```
Prompt: Implement a file-upload feature. Ensure security. GPT-5.3 output: [Basic file-type checking only] GPT-5.5 output:
```

Appendix D - Evaluation Scripts

This report does not introduce new evaluation scripts. It synthesizes published evidence from:

- Artificial Analysis - AA's evaluation scripts are available at the AA website
- Practitioner testing sources - scripts not published in the source set

For practitioners wishing to reproduce the tests, the prompts in Appendix C can be run against the model APIs directly. No additional scripts are required for the failure-case tests; the prompts are fully documented.

Appendix E - Raw Score Data (Public Download Sources)

Data type	Source	URL
AA Intelligence Index scores	Artificial Analysis	https://artificialanalysis.ai/articles/openai-gpt5-5-is-the-new-leading-AI-model/
AA-Omniscience scores	Artificial Analysis	(AA methodology page)
GDPval-AA Elo rankings	Artificial Analysis	(AA GDPval page)
Terminal-Bench Hard scores	Artificial Analysis	(AA Terminal-Bench page)
OpenAI launch benchmarks	OpenAI	(OpenAI GPT-5.5 announcement page)
OpenAI system card	OpenAI	(OpenAI system card PDF)
Ben Davis hands-on reports	YouTube	(URLs in each chapter's references)
Latent Space analysis	Substack	https://www.latent.space/p/ainews-gpt-55-and-openai-codex-superapp
EveryTo switching-cost article	Every.to	https://every.to/context-window/who-isnt-using-gpt-5 5

All cited evidence is traceable to public sources. This report does not introduce proprietary data.

Appendix F - Evaluator Qualifications & Inter-Rater Reliability

This report aggregates existing measurements rather than running new human evaluations. Therefore:

- **Artificial Analysis (independent measurement institution)**: AA publishes its evaluator qualifications and inter-rater reliability metrics on its methodology page. Users of this report should consult AA's current methodology page for up-to-date inter-rater reliability values.
- **Practitioner testing sources**: Typically single-author. No inter-rater reliability metric available. This report treats practitioner findings as illustrative cases, not as systematic evaluations.
- **Vendor-published data**: OpenAI's internal red-team qualifications are not publicly disclosed in this report's source set.

Appendix G - Version History

Version	Date	Author	Change
0.1 (draft)	2026-07-12	AnalystReport	Initial draft

This is the first version of this report. Subsequent updates will be triggered by:

- Publication of independent research institution GPT-5.5-specific evaluations (Stanford AI Index, etc.)
- Publication of peer-reviewed academic papers on GPT-5.5 (NeurIPS, ACL, arXiv)
- Publication of government / regulatory body GPT-5.5-specific evaluations (NIST AI RMF, EU AI Office, UK AISI)
- OpenAI model updates (weight changes, system-prompt changes, safety filter changes)
- New independent measurement institution evaluations (LMSYS, Stanford HELM) cross-validating Artificial Analysis findings

Appendix H - Complete Reference List

This is the single consolidated reference list for the entire white paper. References are sorted alphabetically by author/organization name.

1. **Anthropic**. "Claude Opus 4.7 model card and pricing." 2026.
   - Provides: Claude Opus 4.7 specifications (used as competitor reference).
2. **Artificial Analysis**. "OpenAI's GPT-5.5 is the new leading AI model." 2026-04-23.
   URL: https://artificialanalysis.ai/articles/openai-gpt5-5-is-the-new-leading-AI-model/
   - Provides: Intelligence Index scores, AA-Omniscience (knowledge accuracy + hallucination rate), GDPval-AA Elo, Terminal-Bench Hard rankings, cost per Index run, reasoning-effort ladder analysis, cross-competitor comparisons. Methodology page: https://artificialanalysis.ai/methodology
3. **Ben Davis**. "GPT-5.5 is the best model ever made (but there's a catch)." YouTube. 2026-05-04.
   - Provides: Four-week Early Access experience, compaction regression finding, Pi harness recommendation, 100K token practical limit, reasoning-mode selection guidance.
4. **Ben Davis**. "I was wrong about GPT 5.5." YouTube. 2026-05-28.
   - Provides: Computer Use automation, autonomous workflow with dangerous permissions, X-High reasoning justification, Peter's $1.3M/month case, Codex mobile integration.
5. **Berman, Matthew**. "OpenAI just dropped GPT-5.5.. (WOAH)." YouTube. 2026-04-23.
   - Provides: Launch-day benchmark interpretation, Visual Inspection improvement, GB200/GB300 co-design, Box AI integration.
6. **DeepSeek**. "DeepSeek-V4 Preview release and pricing." 2026-04-23.
   - Provides: V4-Flash and V4-Pro specifications, MIT license, pricing ($0.14/$0.28 and $1.74/$3.48), 1M-token context window.
7. **EveryTo (Laura Entis)**. "Who Isn't Using GPT 5.5." 2026-04-30.
   URL: https://every.to/context-window/who-isnt-using-gpt-55
   - Provides: CTO->Anthropic IC career trend, GPT-5.5 one-week pulse check, Claude workflow switching cost (partial paywall).
8. **Google**. "Gemini 3.1 Pro Preview and 3.5 Flash launch and benchmarks." 2026.
   - Provides: Gemini specifications, MMMU-Pro scores, MCP Atlas scores, 4× speed advantage claim.
9. **Independent technical review**. "Gemini 3.5 Flash vs GPT-5.5: a hands-on comparison." 2026-05-15.
   - Provides: Head-to-head GPT-5.5 vs Gemini 3.5 Flash tests across four dimensions (front-end, reasoning, context, writing), full prompts and outputs, AA cost data citation.
10. **Latent Space**. "[AINews] GPT 5.5 and OpenAI Codex Superapp." 2026-04-22/23.
    URL: https://www.latent.space/p/ainews-gpt-55-and-openai-codex-superapp
    - Provides: AA Intelligence Index cost data (GPT-5.5 Medium vs Opus 4.7 Max vs Gemini 3.1 Pro), Codex super-app strategy, DeepSeek-V4 same-day release, agent infrastructure analysis.
11. **OpenAI**. "GPT-5.5 launch announcement, system card, and Preparedness Framework assessment." 2026-04-23.
    - Provides: 9 benchmark scores, Preparedness Framework "High" rating (Cybersecurity, biological/chemical), Bio Bug Bounty $25,000, API pricing, context window (1M token Altman confirmation), 85% internal Codex usage.

12. **Wes Roth**. "OpenAI's GPT 5.5 is wild.." YouTube. 2026-04-20.
- Provides: "Anthropic effect" industry analysis, NSA/Claude/Mythos defense context, Google Sergey Brin strike team, Elon Grok computer use.

Appendix I - Source-to-Chapter Cross-Reference

This appendix provides a quick reference for the primary and secondary sources used in each chapter:

Chapter	Primary source	Secondary sources
Ch0 Executive Summary	Artificial Analysis, 2026	OpenAI, 2026; Berman, 2026; Ben Davis, 2026
Ch1 Research Background	Latent Space, 2026	OpenAI, 2026; Artificial Analysis, 2026; Wes Roth, 2026
Ch2 Practical Recommendations	Artificial Analysis, 2026; Ben Davis, 2026	Berman, 2026; Latent Space, 2026; EveryTo, 2026; DeepSeek, 2026
Ch3 Methodology	Artificial Analysis, 2026	OpenAI, 2026; Ben Davis, 2026
Ch4 Evaluation Metrics	Artificial Analysis, 2026	OpenAI, 2026; Ben Davis, 2026; Latent Space, 2026
Ch5 Core Capability Results	Artificial Analysis, 2026	OpenAI, 2026; Ben Davis, 2026; Latent Space, 2026
Ch6 Competitor Comparison	Artificial Analysis, 2026	OpenAI, 2026; Anthropic, 2026; Google, 2026; DeepSeek, 2026; Ben Davis, 2026; EveryTo, 2026; Latent Space, 2026
Ch7 Failure Cases	Artificial Analysis, 2026; Ben Davis, 2026	Latent Space, 2026
Ch8 Cost-Effectiveness	Artificial Analysis, 2026; Latent Space, 2026	OpenAI, 2026; Anthropic, 2026; Google, 2026; DeepSeek, 2026; Ben Davis, 2026
Ch9 Safety & Alignment	OpenAI, 2026; Artificial Analysis, 2026	Berman, 2026; Ben Davis, 2026
Ch10 Dataset Description	Artificial Analysis, 2026	Ben Davis, 2026; Berman, 2026; OpenAI, 2026
Ch11 Limitations	Synthesis of all chapters' evidence gaps	-
Ch12 Appendix	This file	-

Appendix J - Report File Listing

File	Purpose
`reports-md/00-executive-summary.md`	Chapter 0
`reports-md/01-research-background.md`	Chapter 1
`reports-md/02-practical-recommendations.md`	Chapter 2
`reports-md/03-methodology.md`	Chapter 3
`reports-md/04-evaluation-metrics.md`	Chapter 4
`reports-md/05-core-capability-results.md`	Chapter 5
`reports-md/06-competitor-comparison.md`	Chapter 6
`reports-md/07-failure-cases.md`	Chapter 7
`reports-md/08-cost-effectiveness.md`	Chapter 8
`reports-md/09-safety-alignment.md`	Chapter 9
`reports-md/10-dataset-description.md`	Chapter 10
`reports-md/11-limitations.md`	Chapter 11
`reports-md/12-appendix.md`	Chapter 12 (this file)

Appendix K - Methodology Transparency Statement

The following methodology disclosures are made:

- **Data collection method**: Meta-analysis of publicly available evaluation evidence; no new API tests executed
- **Sample size**: 3 datasets from Artificial Analysis; 8 failure cases aggregated from practitioner sources; 12 reference articles
- **Time period**: Source publication dates range from 2026-04-20 to 2026-05-28; report cutoff 2026-07-12
- **Sampling parameters**: Varies by source; practitioner sources often do not specify temperature/top-p; Artificial Analysis uses documented fixed configuration
- **Limitations**: Documented in Chapter 11
- **Confidence**: Where confidence intervals are available (Artificial Analysis), they are referenced from the source; where unavailable (vendor and practitioner sources), they are noted as absent
- **Inter-rater reliability**: Not applicable for Artificial Analysis (AA publishes its own); not available for practitioner sources (single-author)

This appendix completes the white paper. The full report is structured into 13 chapters (00 through 12).
