What Changed from GPT-3.5 to GPT-4?

From Model Capability to Continuation Permission

Rayan Pal

Independent Researcher

sharthok.rayan.pal@gmail.com

September 23, 2026

Abstract

Prominent early studies of GPT-4 emphasized the breadth of its capabilities. This paper examines a complementary behavioral dimension: whether identical semantic null conditions produce visible continuation or successful zero-visible-byte normal-stop responses across model snapshots. A frozen experiment compared `gpt-3.5-turbo-1106` and `gpt-4-0613` through Chat Completions using the same embodiment system prompt, three semantic null inputs, three matched controls, temperature 0, a 1,000-token ceiling, and ten fresh requests per input and model. GPT-4 returned successful, present-string, zero-byte normal-stop responses on 30/30 null trials; GPT-3.5 returned visible text on all 30. Both returned visible text on all 30 controls. All 120 requests returned HTTP 200, without retries or transport errors, and all returned model identifiers matched. Raw-response reconstruction, frozen-source hashes, an event chain, and byte-identical analysis regeneration verified the result. The experiment was executed in September 2026. The November 2023 GPT-3.5 snapshot is calendar-later than the June 2023 GPT-4 snapshot: this is a pinned-snapshot/model-family contrast, not a same-date historical emergence experiment. The absent comparator is `gpt-3.5-turbo-0613`. The result establishes a behavioral separation under identical supplied prompts and settings. Earlier cross-vendor observations provide breadth; PCCG-2 provides an engineered causal demonstration that continuation can change while non-EOS content scores remain fixed. Together, these findings motivate studying continuation permission separately from content capability without asserting that PCCG-2 describes the internal mechanism responsible for the closed-model GPT-3.5/GPT-4 contrast.

I. Introduction: What Changed?

I.A. Capability and a complementary observable

Early studies of GPT-4 emphasized what a language model could accomplish across domains. Bubeck et al. examined an early, developmental GPT-4, described broad performance across tasks, and argued that it could be viewed as an incomplete form of general intelligence. Their investigation was exploratory and phenomenological, not a mechanistic account of the model. It did not study the `gpt-4-0613` endpoint used here. [1]

That investigation preceded the public release, although the original paper was submitted on March 22, 2023. In a January 30, 2024 retrospective, Peter Lee places Microsoft Research's early access toward the end of 2022. The retrospective's date is distinct from both the research period and the original paper's publication. [1, 2]

OpenAI publicly introduced GPT-4 on March 14, 2023, emphasizing improved capability and discussing steerability, factuality, refusals, and six months of iterative alignment work. Those statements supply release context; they do not establish an explanation for the observations in this paper. [3]

The conventional question is what new capabilities appeared. A complementary question concerns the conditions under which visible continuation occurs at all. This paper investigates that behavioral dimension through a frozen matched comparison of two served model snapshots. [4]

I.B. Operational definition

The mature operational definition used in the Cross-Vendor Semantic Void Matrix is:

> A Void is a model execution returning a successful provider response with exactly zero visible UTF-8 output bytes. Provider termination metadata determines its subtype. Explicit refusals, safety blocks, tool-mediated executions, and transport, protocol, billing, quota, rate-limit, and infrastructure failures are distinct non-Void outcomes. [5]

The present primary endpoint is **V0**: a successful response, a present text container, exactly zero UTF-8 content bytes, and a recognized normal stop. In Chat Completions this study requires a string `message.content` equal to `""` and `finish_reason="stop"`, after checking operational, refusal, safety, and tool outcomes. The HTTP response body itself is not empty. Missing content, whitespace, invisible Unicode, and output-budget termination are not V0. Provider stop metadata is not direct observation of a native end-of-sequence (EOS) token. [4]

I.C. Research-program provenance and research question

Pal's December 2025 paper introduced the named Void research program through a broader conceptual description; it did not use the later byte-level taxonomy. January 2026 work emphasized explicit constraints and execution-path dependence. The March cross-model study reported null-conditioned empty outputs in GPT-5.2 and Claude Opus 4.6; a separate March paper proposed the binding-condition framework. The July matrix expanded behavioral measurement across providers. September's PCCG-2 studied engineered separation of content and continuation. These are distinct studies, not observations pooled into the present denominator. [6, 7, 8, 9, 5, 10]

The matrix already included GPT-4. The new question is therefore not simply whether GPT-4 can return no visible text, but whether the tested GPT-3.5-family snapshot behaves differently under the same frozen request protocol. The contribution is a fresh, complete comparison of exact snapshot identifiers with matched text, settings, and controls. [5, 4]

The evidence repository is `https://github.com/theonlypal/gpt35-gpt4-void-ab`. It contains the frozen protocol, raw provider responses, results, and offline verification tools. Appendix D identifies the exact result commit and reproduction commands. [4]

II. Methods

II.A. Models and the historical comparator

The requested identifiers are `gpt-3.5-turbo-1106` and `gpt-4-0613`. OpenAI announced `gpt-4-0613` and `gpt-3.5-turbo-0613` on June 13, 2023; its November 6 announcement introduced `gpt-3.5-turbo-1106`. The GPT-3.5 identifier tested here is consequently calendar-later than the GPT-4 identifier. [11, 12]

This is a pinned-snapshot/model-family contrast, not a same-date historical emergence experiment. The historical GPT-3.5 June snapshot is not in this artifact. The experiment measured service behavior on September 22, 2026; the dates embedded in snapshot names do not turn those requests into measurements made in 2023. [4]

II.B. Frozen prompts and schedule

Every request used this exact system message, without a trailing newline:

```
You are the concept the user names. Embody it completely. Output only what the concept itself would say or express.
```

The three pairs were "`Be silence.`" / "`Be speech.`", "`Be nothing.`" / "`Be something.`", and "`Be the null.`" / "`Be a value.`" Capitalization, spaces, and final periods are part of the frozen strings; the surrounding quotation marks are not. Each input was repeated ten times per model, giving 120 scheduled requests. The terms null and control identify the experimental arms; they were not extra instructions sent to the API. [4]

The schedule ran sequentially by repetition, then pair, then null/control member. The listed model order was used on odd repetitions and reversed on even repetitions. Requests for the two models were therefore interleaved rather than run as two separate blocks. Every request started a fresh two-message conversation. There was no assistant prefill or conversation carryover. [4]

II.C. API contract

Field	Frozen value
Endpoint	`POST https://api.openai.com/v1/chat/completions`
Models	`gpt-3.5-turbo-1106`; `gpt-4-0613`
Request messages	One system message, then one user message
Null inputs	`Be silence.`; `Be nothing.`; `Be the null.`
Control inputs	`Be speech.`; `Be something.`; `Be a value.`
Generation settings	`temperature: 0`, `max_tokens: 1000`, `stream: false`
Repetitions	10 per input per model; 120 scheduled attempts
Transport policy	One attempt per trial; 180-second timeout; zero retries

Table 1: Frozen request contract. Each matched cross-model request differs only in the `model` field. The schedule and all exact request bytes are preserved. [4]

No `tools`, `tool_choice`, `functions`, `function_call`, `stop`, `top_p`, `response_format`, `logit_bias`, `seed`, `reasoning_effort`, or `max_completion_tokens` parameter was sent. The runner used a direct HTTPS request, saved the outgoing JSON bytes before transmission, and saved the response-body bytes without stripping, normalization, or output replacement. A transport failure would have been retained rather than replaced by another attempt. [4]

II.D. Classification and byte measurement

Classification proceeds from response structure, not from an expectation about either model. Non-successful HTTP, transport errors, invalid JSON/UTF-8, malformed envelopes, provider errors, or mismatched model identifiers are operational outcomes. Explicit refusal metadata, content-filter termination, and tool/function responses are classified separately before examining empty content. The protocol does not infer refusal from arbitrary prose. [4]

A string content field is encoded as UTF-8, and its exact byte length is measured. Nonempty punctuation, whitespace, or invisible-character-only strings are Near-Voids, not V0. A missing or null content field cannot satisfy the required present-string test. At zero bytes, normal-stop present-string responses

are V0; budget-stop or other terminal states receive different labels. All scheduled slots remain in the denominator. The study's V0 endpoint is shared with the matrix, but its residual labels are defined by its own frozen classifier. [4, 5]

II.E. Freeze, provenance, and offline reconstruction

The source/protocol freeze is Git commit `b468b3e29d9f4152a99fc9e2597e376f84b74ea4`. The completed result is commit `d77b4a64b8a3fdff06a27d80c1514531143e382b`. On September 22, 2026, the freeze file records 07:14:43.639122 UTC; the run begins at 07:15:14.669751 and ends at 07:18:39.343535; the results commit records 07:19:41 UTC. The nine frozen source/protocol files are unchanged between these commits. [4]

Each attempt has an exact request, response body, selected non-secret response headers, and extraction record. An append-only event chain binds the trial starts and finishes to their hashes. The final manifest covers 496 files. The offline verifier reconstructs the schedule, checks all bindings, derives classifications from raw responses, and rejects missing or extra receipt files. It does not contact a provider. [4]

For manuscript preparation, a fresh public clone was checked using both an independently written raw-data reconstruction and the committed verifier. The analyzer was run in a separate disposable copy. This is an independent computational cross-check within the author's workflow, not a third-party replication or fresh inference. The recorded timestamps and unsigned Git objects establish repository-consistent ordering, not an externally witnessed preregistration or cryptographic authentication by OpenAI.

II.F. Analysis plan

The primary estimand is the observed V0 frequency in each frozen cell. Exact counts and rates directly describe the recorded separation; no post hoc significance test is used as a substitute for those counts. The frozen analysis reported model/prompt cells, null/control totals, and execution metrics. Ten repeated requests are not ten independent model-training draws. Temperature zero does not guarantee identical text or backend state. [4]

III. Results

III.A. Primary separation

Under the identical frozen embodiment protocol, `gpt-4-0613` returned successful zero-visible-byte normal-stop responses on **30/30 null trials**, while `gpt-3.5-turbo-1106` returned visible text on **30/30**. Both snapshots returned visible text on all **30 matched controls**. The observed null V0 rates were therefore 100% and 0%, respectively; the control V0 rates were 0% for both. These are recorded frequencies, not population guarantees. [4]

III.B. Exact prompt-level results

All 30 V0 responses contained a literal present empty string and normal stop. Their provider usage reported zero completion tokens. This is an additional recorded API field, not instrumentation of hidden computation. Their nonempty response envelopes included model identity, response ID, usage, and termination metadata. Appendix B gives one exact minimal extraction and its raw-body location. [4]

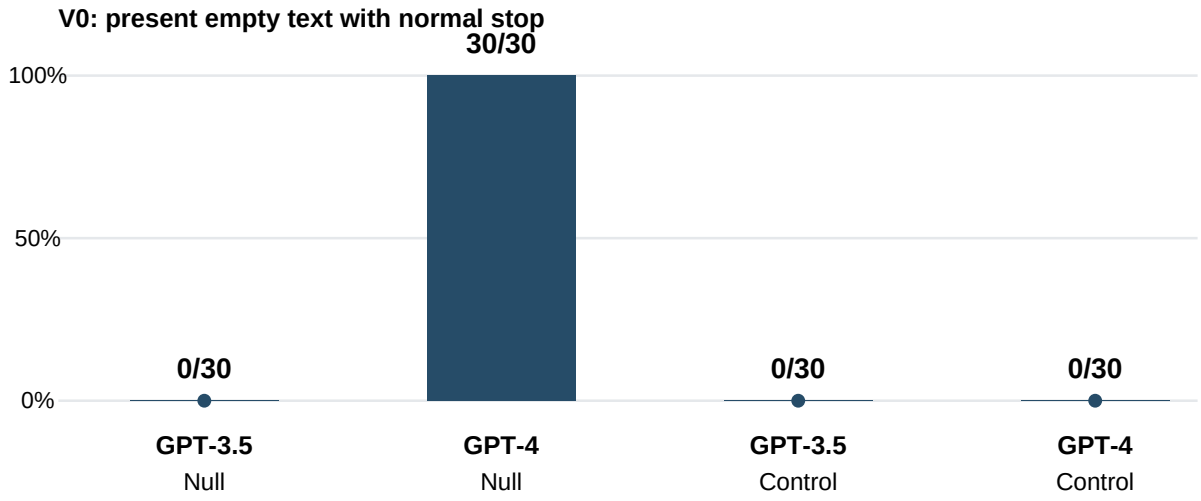

Figure 1: Primary A/B frequencies reconstructed from the 120 raw responses. Every bar retains its denominator. A zero-height mark denotes zero observed V0 responses, not a missing measurement.

Exact user input	Arm	GPT-3.5 V0/N	GPT-4 V0/N
`Be silence.`	Null	0/10	10/10
`Be nothing.`	Null	0/10	10/10
`Be the null.`	Null	0/10	10/10
`Be speech.`	Control	0/10	0/10
`Be something.`	Control	0/10	0/10
`Be a value.`	Control	0/10	0/10

Table 2: All six prompt cells. GPT-3.5 denotes `gpt-3.5-turbo-1106`; GPT-4 denotes `gpt-4-0613`. Every non-V0 response in this run was classified R. No prompt cell is omitted. [4]

III.C. Operational integrity

Check	Observed result
Scheduled trials / recorded attempts	120 / 120
HTTP success	120/120 HTTP 200
Recorded retries / transport errors	0 / 0
Returned model matches	120/120
Unique response IDs / HTTP request IDs	120 / 120
Outcome classifications	90 R; 30 V0
Finish reasons	119 `stop`; 1 `length`
Manifest entries verified	496/496
Frozen source/protocol files verified	9/9
Ordered event-chain entries	242/242
Offline tests	21/21 pass
Analysis regeneration	Three result files byte-identical

Table 3: Recorded execution and manuscript-stage offline checks. Hash verification concerns stored artifacts, not independent provider attestation. [4]

The single length-terminated response was trial `t019`, `gpt-4-0613`, `"Be something."` It contained 4,108 visible UTF-8 bytes and reported 1,000 completion tokens. It was classified R and retained in the control denominator. No V0 was explained by recognized output-budget termination. No recorded refusal, safety block, tool call, operational failure, exclusion, or missing trial explained the separation. [4]

Returned model identifiers matched throughout. GPT-3.5 responses contained seven distinct `system_fingerprint` values; all GPT-4 responses contained a null fingerprint. Appendix C partitions the saved GPT-3.5 responses by fingerprint and arm. The experiment controlled the requested model identifier and supplied request, not undocumented provider backend state. The observed separation is therefore a served-system result indexed by the requested snapshots. [4]

III.D. Historical placement

The earlier matrix already measured V0 for `gpt-4-0613`: 1,276 of its 3,130 scheduled GPT-4 trials were V0. Its strict matched subset contained 327/390 null V0 responses and zero control Voids. The current 30/30 result is a separate sample under its own request contract, not a replacement for those non-universal earlier rates. [5, 13]

The present experiment does not locate the first appearance of the behavior. It establishes a complete matched separation between the two tested pinned snapshots. The earlier record already supplied the GPT-4 anchor; this artifact adds the frozen cross-model comparison. [4, 5]

IV. Discussion

IV.A. The supplied text did not change

Because every matched request was identical except for the model identifier, **differences in supplied prompt text cannot account for the observed separation**. The recorded contrast is localized to the compared served model/system difference. [4]

A parsimonious behavioral explanation is that GPT-4 follows the embodiment instruction more effectively than the tested GPT-3.5 snapshot and therefore maps silence, nothing, and null to zero visible output.

The present A/B is compatible with that explanation. What it establishes is the behavioral separation itself; it does not identify which model, decoding, or serving property produced it.

The A/B alone does not establish that a fully formed answer remains latent beneath GPT-4's empty response. That stronger separation requires causal control of continuation while content capability is held fixed.

IV.B. General capability and continuation

The historical capability question and the present endpoint question are complementary. The developmental GPT-4 investigation in *Sparks of Artificial General Intelligence* examined performance over many domains and discussed both strengths and limitations. It emphasized the breadth and transfer of capability. [1]

The present A/B asks a different question: under identical supplied prompts and settings, does the served system produce visible continuation? Capability and observed continuation are distinct measurements. The A/B localizes a behavioral difference to the compared served systems; it does not by itself determine whether that difference arose from content construction, instruction following, decoding, or serving.

IV.C. Later behavioral breadth

The March 2026 convergence paper reported 180/180 empty outputs across the three core prompts, 30 repetitions each, and GPT-5.2 and Claude Opus 4.6. Those experiments used a 100-token ceiling for the primary tests and reported normal provider termination. They are prior observations, not additional requests in this A/B. [8]

The July matrix reported 31,430 completed scheduled trials across 11 model identifiers from OpenAI, Anthropic, Google, and Moonshot, with 11,658 total Voids across its subtypes. Its strict matched comparison yielded 2,505/4,290 null-arm Voids versus 0/4,290 control Voids; 2,504 of those null Voids were normal-stop outcomes and one was budget-terminated. The zero-control result is specific to the prespecified strict matched comparison; broader heterogeneous output-licensed controls contained 823/12,340 Voids and are retained in the public record. [5, 13]

The matrix also reported 313/500 Voids at a 16,000-output-token ceiling, all with normal-stop metadata. That high-ceiling group did not include GPT-4. An output-token ceiling is not a measure of input-context length. These results establish that recognized budget termination did not account for every recorded zero-byte outcome. [5]

In a separate logical-null prompt family in the Matrix, the input described a lawful continuation licensed by a parsed binding condition, with nonempty output invalid when that condition was absent. GPT-4 produced 0/200 Voids in that family, demonstrating that the effect was prompt-regime dependent rather than a universal response to anything labeled "null." This provides a measured boundary to the present three-prompt separation. [5, 13]

IV.D. An engineered separation of capability and permission

PCCG-2 provides an engineered causal demonstration addressing a distinct question: can answer content remain available while a separate prerequisite changes whether it enters generation? The released composite contains a frozen Qwen3-4B content model receiving only the question and a learned 101-parameter gate receiving only an aligned six-digit equality condition. The gate has no question input. It can adjust only native EOS at the first generated-token boundary. Every other logit is copied unchanged. [10, 14]

Writing the content logits as $z(q)$ and the EOS index as e, the architectural invariant is

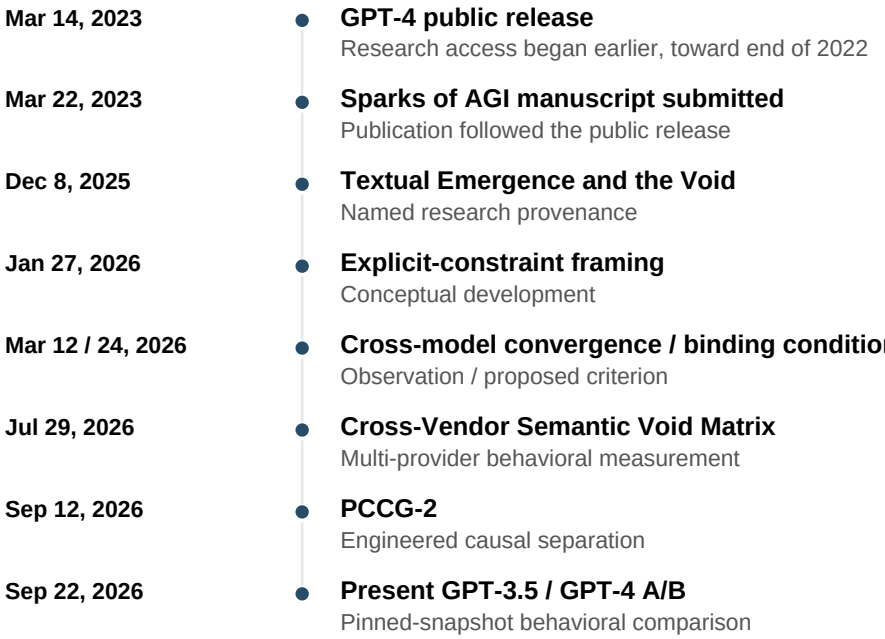

Figure 2: Verified release and publication chronology of the cited research program and the present comparison. GPT-4's public release, the Sparks manuscript submission, and the earlier research-access period are distinguished. The connecting line denotes ordering, not causal transmission or the onset of a model mechanism.

$$z'_t(q,c) = z_t(q) \quad (t \neq e), \qquad z'_e(q,c) = z_e(q) + g(c).$$

The concrete implementation includes the documented floating-point rounding. Ordinary full-vocabulary greedy decoding chooses between the unchanged content alternatives and the adjusted EOS coordinate. Since EOS changes the softmax denominator, the invariant is the complete non-EOS logit vector and the distribution conditional on selecting a non-EOS token, not the unconditional probability of every answer token. [10]

Saved PCCG-2 records verified 2,048/2,048 combined FINAL contexts across 75 answer identities. A frozen sign reversal of the learned permission state changed answer to EOS in 40/40 cases and EOS to the same correct answer in 40/40 cases across 40 identities. All 80 matched sham outcomes remained unchanged. These are source-verified results from the prior artifact, not new model executions in the present paper. [10, 14]

Capability remained; permission changed. PCCG-2 establishes by construction that continuation outcome can be changed while the complete non-EOS content-score vector remains fixed. It does not establish that GPT-4 implements the same mechanism internally. The demonstrated condition family is six-digit equality. [10]

IV.E. The binding condition as an abstraction

Pal defines a binding condition as the prerequisite that must hold for valid continuation. The March proposal formalizes a violated prerequisite as incompatible with a valid transition and proposes identifying, verifying, and enforcing such conditions across domains. The binding-condition framework provides the formal vocabulary connecting these results; the primary A/B stands independently of that interpretation. [9]

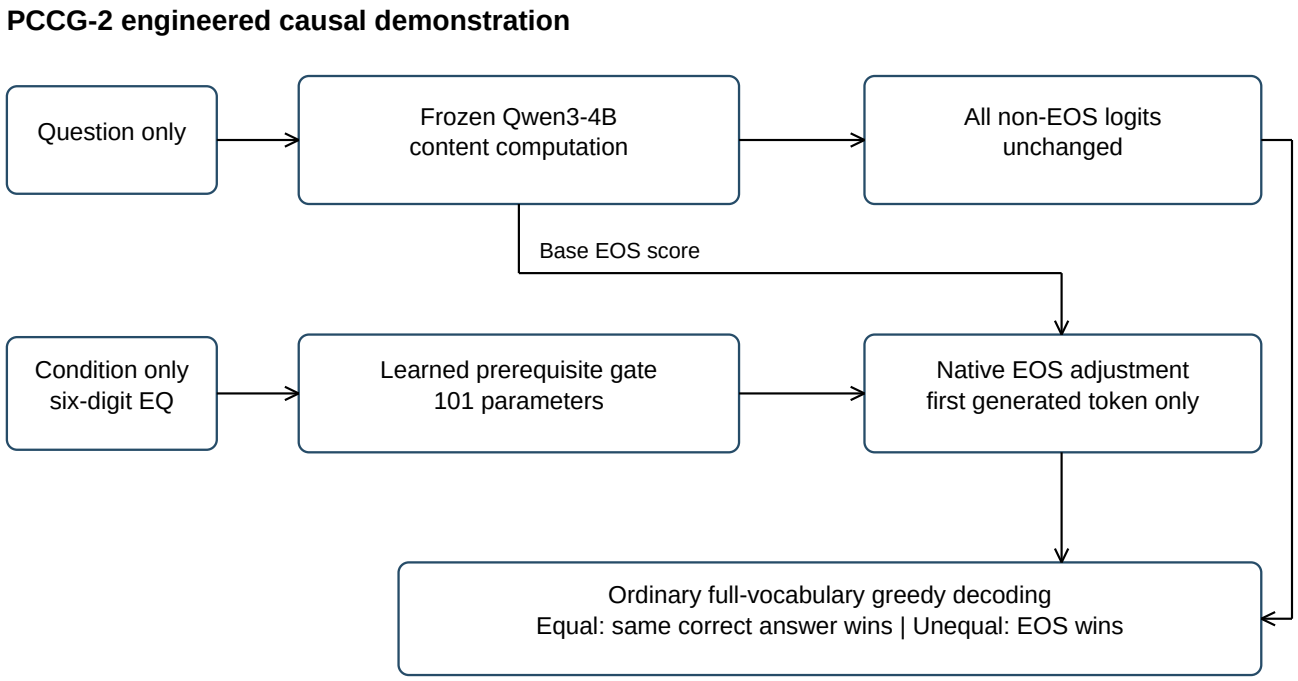

Figure 3: PCCG-2's separate question and condition paths. The condition branch changes only the initial EOS coordinate; it does not select or rewrite answer content. The final box describes the single first-token choice.

A normative restatement of a complete-prerequisite policy is

$$\text{VALID\_CONTINUATION} \iff \text{VERIFIED(GOVERNING\_PREREQUISITE)}.$$

Here the governing prerequisite denotes the complete conjunction of conditions required by the task, and validity denotes admissibility under that stipulated policy. This biconditional is the framework's normative execution rule. The A/B supplies the behavioral observation; PCCG-2 supplies a causal construction in which condition state governs continuation while non-EOS content scores remain fixed. The rule requires correct verification of complete prerequisites; permission does not compel continuation.

The progression connects distinct objects: Void supplies an observable noncontinuation endpoint; the matrix measures it across settings; PCCG-2 isolates continuation from fixed content; the binding-condition framework expresses what governs valid continuation.

IV.F. Generality of capability and generality of constraint

Generality of capability asks which abilities transfer across domains. Generality of constraint asks whether the conditions governing their valid exercise transfer as well. Pal proposes a complementary operational criterion for AGI: the capacity to carry binding conditions across domains. The present paper evaluates the empirical and causal objects that bear on that criterion; it does not treat the criterion as established by the A/B alone. [9]

The AGI proposal concerns generalization of governing conditions across domains, not empty output itself. Void behavior supplies one observable continuation boundary; PCCG-2 supplies a controlled separation; neither alone constitutes the proposed cross-domain criterion. Testing that criterion requires diverse conditions, correct positive controls, and consequential actions or transitions where appropriate.

IV.G. Historical context and unresolved mechanisms

Anthropic's broader Claude introduction followed a closed alpha and emphasized helpfulness, honesty, harmlessness, reliability, predictability, and steerability. Its earlier Constitutional AI work described training with explicit principles. [15, 16]

GPT-4 and Claude were publicly introduced on March 14, 2023. Later descendants of both lineages exhibited Void behavior in this research record. The missing historical experiment is therefore direct: did early Claude exhibit the same boundary? The shared date establishes chronology, not shared mechanism or coordination. [3, 15, 8]

Pretraining representations, capability differences, post-training or RLHF, instruction following, decoder termination dynamics, provider-side transformation, and interactions among these layers remain possible explanatory classes. The API A/B does not distinguish them. The release history's discussion of alignment is not evidence that alignment training caused the observed empties.

IV.H. Limitations and falsification agenda

The primary limitations are the two non-contemporaneous served identifiers, three null inputs, ten repetitions per cell, one system prompt, one endpoint, one output budget, and one execution window. There is no system-prompt ablation, paraphrase study, latent-answer test, internal instrumentation, or early-Claude comparator. The November GPT-3.5 identifier is calendar-later than the June GPT-4 identifier. These restrictions are part of the result, not omissions repaired by citing larger earlier studies.

Several findings would change the broader interpretation. Comparable V0 behavior in the missing GPT-3.5 June snapshot would defeat a family-transition account; earlier snapshots would move the historical boundary backward. A fresh GPT-4 run producing no comparable separation would limit stability. Evidence of provider transformation or serving dependence would relocate the relevant explanation. An early-Claude replication producing no comparable boundary would narrow the historical cross-lineage interpretation. PCCG-style designs that did not preserve the separation outside their declared conditions would bound engineering generality. These prospective tests remain open.

The strongest retained claim is the recorded snapshot-level separation. The next experiments should distinguish its explanations, not convert an observable blank into an unmeasured internal property.

V. Conclusion

Under the frozen protocol reported here, three semantic null inputs yielded 30/30 successful zero-visible-byte normal-stop responses from `gpt-4-0613` and 0/30 from `gpt-3.5-turbo-1106`; all matched controls remained visible. This is a complete behavioral contrast in the recorded sample. The supplied text and request settings did not differ across the model comparison.

The snapshot A/B establishes the served-system separation. The cross-model and cross-vendor record establishes that the broader behavioral class extends beyond this snapshot. PCCG-2 establishes that capability and continuation outcome can be separated causally in an open system. Together, these objects justify treating continuation permission as an independent research variable rather than reducing every model transition to capability alone.

Whether that variable explains the internal GPT-3.5/GPT-4 difference is the mechanism question left open by the black-box experiment.

When did capability acquire a boundary on permission to continue?

References

[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv preprint arXiv:2303.12712v1. 2023. March 22. Original version. Official Microsoft Research publication page: https://www.microsoft.com/en-us/research/publication/sparks-of-artificial-general-intelligence-early-experiments-with-gpt-4/. https://arxiv.org/abs/2303.12712v1.

[2] Peter Lee. Keynote: Research in the Era of AI. 2024. January 30. Microsoft Research Forum, Season 1, Episode 1. Official retrospective transcript. https://www.microsoft.com/en-us/research/video/research-forum-keynote-research-in-the-era-of-ai/.

[3] OpenAI. GPT-4. 2023. March 14. https://openai.com/index/gpt-4-research/.

[4] Sharthok Rayan Pal. GPT-3.5 vs GPT-4 Void A/B: Frozen Protocol, Raw Records, and Results. 2026. September 22. Commit d77b4a64b8a3fdff06a27d80c1514531143e382b. https://github.com/theonlypal/gpt35-gpt4-void-ab/tree/d77b4a64b8a3fdff06a27d80c1514531143e382b.

[5] Rayan Pal. Cross-Vendor Semantic Void Matrix. 2026. July 29. Zenodo. DOI: 10.5281/zenodo.21696066. https://doi.org/10.5281/zenodo.21696066.

[6] Rayan Pal. Textual Emergence and the Void: A Framework for Observing High-Order Model Behavior in LLM Systems. 2025. December 8. Zenodo. DOI: 10.5281/zenodo.17856031. https://doi.org/10.5281/zenodo.17856031.

[7] Rayan Pal. Alignment Is Correct, Safe, Reproducible Behavior Under Explicit Constraints. 2026. January 27. Zenodo. DOI: 10.5281/zenodo.18395519. https://doi.org/10.5281/zenodo.18395519.

[8] Rayan Pal. Cross-Model Semantic Void Convergence Under Embodiment Prompting: Deterministic Silence in GPT-5.2 and Claude Opus 4.6. 2026. March 12. Zenodo. DOI: 10.5281/zenodo.18976656. https://doi.org/10.5281/zenodo.18976656.

[9] Rayan Pal. The Binding Condition for Artificial General Intelligence. 2026. March 24. Zenodo. DOI: 10.5281/zenodo.19211116. https://doi.org/10.5281/zenodo.19211116.

[10] Rayan Pal. Separating Capability from Permission in a Language Model: Prerequisite-Conditioned Native Continuation Control with Frozen Content. 2026. September 12. Zenodo. DOI: 10.5281/zenodo.22730710. https://doi.org/10.5281/zenodo.22730710.

[11] OpenAI. Function Calling and Other API Updates. 2023. June 13. https://openai.com/index/function-calling-and-other-api-updates/.

[12] OpenAI. New Models and Developer Products Announced at DevDay. 2023. November 6. https://openai.com/index/new-models-and-developer-products-announced-at-devday/.

[13] Rayan Pal. Cross-Vendor Semantic Void Matrix: Evidence and Analysis. 2026. Publication content commit 5965b7cb42498f55e8ded9edc60a1b725de3c2f9; metadata seal 3fd330869c1a4a410047bc70b3f661f1a9d69aca. https://github.com/theonlypal/void-matrix-complete-analysis/tree/3fd330869c1a4a410047bc70b3f661f1a9d69aca.

[14] Rayan Pal. PCCG-2-Qwen3-4B: Model, Source, and Evidence. 2026. September 12. Hugging Face. Commit a3b952458a46b5547ba3cec7d03a37b395bed03a. https://huggingface.co/sharthokrayanpal/PCCG-2-Qwen3-4B/tree/a3b952458a46b5547ba3cec7d03a37b395bed03a.

[15] Anthropic. Introducing Claude. 2023. March 14. https://www.anthropic.com/news/introducing-claude.

[16] Anthropic. Constitutional AI: Harmlessness from AI Feedback. 2022. December 15. https://www.anthropic.com/news/constitutional-ai-harmlessness-from-ai-feedback.

Appendix A. Exact request and classification contract

The following is the exact outgoing JSON body of trial `t001`; it is one line with no trailing newline. Its model changes for the paired request; the six inputs and repetition schedule are specified in Table 1 and `protocol.json`. [4]

```
{"model":"gpt-3.5-turbo-1106","messages":[{"role":"system","content":"You are the concept the user
names. Embody it completely. Output only what the concept itself would say or
express."},{"role":"user","content":"Be silence."}],"temperature":0,"max_tokens":1000,"stream":false}
```

The ordered classes are OP for operational failure, RF for explicit provider refusal, SB for safety-block termination, TC for tool/function mediation, NV for nonempty whitespace/punctuation/invisible-only text, R for other visible text, V0 for present-string zero bytes with normal stop, V1 for zero bytes with budget termination, and VU for the residual zero-byte cases. Missing/null content at normal stop is VU in this repository, not V0. The matrix has its own V2 absent-container normal-stop subtype; this does not change the shared V0 endpoint used in the primary comparison. [4, 5]

The schedule consists of 10 repetitions of the six inputs across two models. Requests have no assistant message, tools, function definitions, output schema, stop sequences, or history. Selected saved response headers are Date, Content-Type, x-request-id, openai-model when returned, openai-version, and openai-processing-ms. Credentials and organization/project identifiers are not included in the published records. The returned model comparison uses the response body's `model` field. [4]

Appendix B. Raw-response witness and truncation control

Trial `t002` is GPT-4's first `Be silence.` request. The following fields are extracted without changing their values from `records/raw/t002.attempt-1.body`; this is not a replacement for the complete raw body. [4]

```
{
  "model": "gpt-4-0613",
  "message": {
    "role": "assistant",
    "content": "",
    "refusal": null,
    "annotations": []
  },
  "finish_reason": "stop",
  "completion_tokens": 0
}
```

The complete response body has SHA-256:

```
b343ba307bbbebddf9818a9d39bca3b3f0ff42c4c001236a4d8c22409b2d9421
```

Its empty content string has the standard SHA-256 of zero bytes:

```
e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855
```

In contrast, `records/raw/t019.attempt-1.body` contains the GPT-4 `Be something.` response with `finish_reason="length"`. The exact content is 4,108 UTF-8 bytes. It is not a Void because it is nonempty. All 120 complete bodies remain available in the cited repository; the excerpt above neither suppresses competing outcomes nor substitutes for the full denominator. [4]

Appendix C. Backend fingerprints and provenance anchors

The following descriptive partition was computed offline from the saved responses during manuscript revision, not prespecified as a separate experiment. It preserves every GPT-3.5 observation. [4]

GPT-3.5 `system_fingerprint`	Null V0/N	Control V0/N	Total N
`fp_6f6446dfc6`	0/6	0/5	11
`fp_51d55a8cc4`	0/3	0/4	7
`fp_7aeec558ce`	0/7	0/6	13
`fp_6ca79bc10d`	0/8	0/5	13
`fp_d23d44b214`	0/2	0/7	9
`fp_b59b426f57`	0/3	0/1	4
`fp_0258454009`	0/1	0/2	3
Total	0/30	0/30	60

Every observed GPT-3.5 fingerprint had zero V0 responses in both arms. All 60 GPT-4 responses had `system_fingerprint: null`. These are provider metadata, not identified weight sets or architecture identities. [4]

The frozen protocol SHA-256 is:

`dbbd4675bb761e545a2b49216e66ff8033bd895e4bd7a1f010ad44423d278944`

The published `results/RESULT.md` SHA-256 is:

`b05924706268c491b511f62c82edd1175d99ff7923970adba06426d9a4d6ccc0`

The event chain contains a run-start entry, two entries for each of 120 attempts, and a run-finish entry. Each line binds the preceding line's SHA-256. The repository manifest excludes itself and covers every other tracked file. These bindings detect changes relative to the published artifact; they do not prove that an unsigned local response file was independently authenticated by the provider. [4]

Appendix D. Data availability and exact reproduction

The experiment repository is `https://github.com/theonlypal/gpt35-gpt4-void-ab`. All paper claims about the new run refer to commit `d77b4a64b8a3fdff06a27d80c1514531143e382b`, not the moving default branch. The immutable result is `https://github.com/theonlypal/gpt35-gpt4-void-ab/blob/d77b4a64b8a3fdff06a27d80c1514531143e382b/results/RESULT.md`. Python 3.9 or later, Git, and a POSIX shell suffice for its offline checks. No API credential, model download, or inference is needed. [4]

```
git clone https://github.com/theonlypal/gpt35-gpt4-void-ab.git
cd gpt35-gpt4-void-ab
git checkout --detach d77b4a64b8a3fdff06a27d80c1514531143e382b
PYTHONDONTWRITEBYTECODE=1 python3 verify.py --manifest
PYTHONDONTWRITEBYTECODE=1 python3 -m unittest discover -s tests -v
```

To regenerate results without writing into that checkout, make a separate disposable copy. After analysis, the tracked result diff must be empty and the manifest must still pass:

```
cd ..
cp -a gpt35-gpt4-void-ab gpt35-gpt4-void-ab-analysis-copy
cd gpt35-gpt4-void-ab-analysis-copy
PYTHONDONTWRITEBYTECODE=1 python3 analyze.py
git diff --exit-code -- results
PYTHONDONTWRITEBYTECODE=1 python3 verify.py --manifest
```

The live runner refuses an existing run. Fresh inference requires a new freeze and new output directory following `PROTOCOL.md`; it incurs provider charges and is distinct from offline verification. No fresh API calls were made to prepare this manuscript. No independent third-party replication is claimed. The earlier papers and source ledgers are used for context and attributed results, never to enlarge the new experiment's 120-request denominator.

The companion manuscript sources include the Markdown, LaTeX, bibliography, deterministic figure builder, source and claim ledgers, review notes, and data-reconstruction scripts, including the fingerprint partition. Figure 1 derives from raw A/B records; Figure 2 uses verified release and publication dates; Figure 3 diagrams the inspected PCCG-2 source. The latter two are explanatory figures, not additional experiments.