ChatGPT

# Ocena naukowa tezy i planu testów amortyzacji kolapsu Cloud+AI

## Działania na repozytoriach i dostępne artefakty

Badanie zgodnie z wymaganiem rozpocząłem od użycia włączonego konektora do entity["company","GitHub","code hosting platform"] i pracy **wyłącznie** na repozytoriach: `DonkeyJJLove/sbom`, `DonkeyJJLove/writeups`, `DonkeyJJLove/chunk-chunk`, `DonkeyJJLove/swarm`, `DonkeyJJLove/ai_platform`, `DonkeyJJLove/HA2D`. Wykonałem operacje: (a) pobranie metadanych repo (operacja typu "get_repo"), (b) pobranie treści kluczowych plików (operacje typu "fetch"), (c) próby wyszukiwania symboli/łańcuchów w repo (operacje typu "search"; w tym środowisku zwracały wynik pusty dla części repo, co traktuję jako ograniczenie indeksu, a nie dowód braku tekstu).

Istotne ograniczenie techniczne w kontekście „absolutnego reżimu naukowego": konektor GitHub w tym środowisku zwraca treść plików jako tekst bez markerów liniowych `filecite`. W konsekwencji **nie mogę cytować linia-po-linii plików repo w formacie filecite**, mimo że pliki zostały pobrane. W raporcie podaję więc (1) ścieżki i funkcję każdego artefaktu, (2) minimalne fragmenty treści jako "dowód z artefaktu", oraz (3) cytowania webowe dla kryteriów metodologicznych, teorii kaskad, FinOps i constraintów energetycznych (tam, gdzie wnioski dotyczą świata poza repo).

Pobrane i przeanalizowane artefakty (rdzeń dowodowy) są następujące:

| Repo | Pobrane pliki | Dlaczego te pliki są krytyczne dla oceny naukowej |
|---|---|---|
| `sbom` | `lab/jenkins/pipeline_one.pipeline`, `docs/03_JENKINS_PIPELINE.md` | To "źródło prawdy" procesu: generacja SBOM → scan → snapshot → delta → gate → event store. Umożliwia testy: deterministyczność delty, false positive/negative gate, replikowalność. |
| `swarm` | `infrastructure/istio/policies/circuit-breaker.yaml`, `infrastructure/istio/policies/rate-limit.yaml` | To implementacja progów runtime, które mają zatrzymywać kaskady (outlier detection, rate limiting). Weryfikowalne testami obciążeniowymi i kaskadowości. |
| `HA2D` | `context_protocol.md` | Definicja protokołu pamięci kontekstu (UUID, timestamp, payload, sha256) i operacji STORE/RETRIEVE, czyli baza replikowalności i dowodowości kontekstu. |

| Repo | Pobrane pliki | Dlaczego te pliki są krytyczne dla oceny naukowej |
|---|---|---|
| `chunk-chunk` | `hmk9d_protocol.yaml` | Formalny kontrakt decyzyjny (S, Σ, A, F, g, H, a*), osie 9D, bramki i invarianty – "teoria sterowania" dla Twojej tezy. |
| `ai_platform` | `platform.md` | Mapowanie woluminu QV9D ↔ struktura katalogów i artefaktów (SPEC/STATE/METRICS/RITUAL/CI) + jawne miejsce "do zaprojektowania" dla deterministycznego ID. |
| `writeups` | `protokoly_kontekstu_chunk-chunk_facebook_case.md` | Metodyka: definicje, "dowód warunkowy i falsyfikacja", ryzyka przeuczenia, dryfu i złej bazy – to bezpośrednio wspiera ocenę jakości naukowej testów. |

Dla kontrapunktu "rzeczywistości" (constraintów poza kodem) przyjmuję jako twarde źródło scenariuszowe analizę entity["organization","International Energy Agency","energy agency"] o energii i AI, w tym o tempie budowy data center (2–3 lata) vs. dłuższych lead-time'ach infrastruktury energetycznej i o projekcji globalnego zużycia energii przez data center do ok. 945 TWh w 2030 w scenariuszu bazowym. [1]

## Pytania badawcze i operacjonalizacja tezy

W "najostrzejszym" sensie naukowym teza jest wartościowa tylko wtedy, gdy da się ją obalić. Dlatego ocena nie zaczyna się od retoryki "czy to brzmi", lecz od pytań, które muszą zostać rozstrzygnięte, żeby wniosek mógł mieć status naukowy (w sensie falsyfikowalności, o której entity["people","Karl Popper","philosopher of science"] pisał jako kryterium demarkacji). [2]

Kluczowe pytania badawcze (z operacjonalizacją):

Czy podejście przechodzi z "seat economics" na "work-unit economics"?
Co mierzymy: koszt na jednostkę pracy (WU) i jego rozkład (medianę i ogon) zamiast kosztu na użytkownika. Jak: korelacją metryk zużycia (czas, requesty, tokeny/compute proxy) z identyfikatorem WU w telemetryce. Dlaczego krytyczne: w AI koszty rosną z wolumenem iteracji/agentów, a nie liniowo z liczbą ludzi; brak meteringu w jednostkach pracy czyni pricing "sygnałem alarmowym", nie regulatorem, co jest rdzeniem Twojej tezy i spójne z ramą FinOps (praktyka zarządzania wydatkami technologicznymi jako proces danych i współpracy). [3]

Czy gating i progi runtime realnie zapobiegają kaskadom?
Co mierzymy: "kaskadowość" (propagacja błędów downstream), MTTR, oraz zmianę reżimu degradacji (np. 429 zamiast 5xx) pod obciążeniem. Jak: testy burst i sustained + fault injection. Dlaczego krytyczne: modele progowe i kaskady opisują, że mały impuls może wywołać rzadkie, ale duże "global cascades", zależnie od progów i struktury sieci. [4]

Czy kontekst jest dowodowy, replikowalny i wersjonowalny?
Co mierzymy: integralność (hash), deterministyczność serializacji, zdolność do wyjaśnienia rozbieżnych wyników różnicą kontekstu (delta-kontekstu), a nie "szumem". Jak: testy STORE/RETRIEVE + "replay runs" + analiza driftu. Dlaczego krytyczne: bez protokołu kontekstu system staje się niefalsyfikowalny operacyjnie (nie ma "zdań bazowych" w sensie Poppera, tylko narrację). [2]

Czy protokół 9D i mapowanie QV9D→artefakty tworzą mierzalny mechanizm sterowania, czy tylko język?
Co mierzymy: czy osie/stany/progi generują wymuszalne bramki (CI/runtime) oraz spójne metryki raportowane per moduł i per WU. Jak: sprawdzenie, czy "invariants/gates" istnieją jako testy automatyczne z warunkami STOP/GO i czy mają ślad w telemetryce. Dlaczego krytyczne: to decyduje, czy teza jest naukowym modelem sterowania, czy literackim opisem.

Czy ekonomicznie następuje amortyzacja ogona ryzyka?
Co mierzymy: redukcję EV(straty) i redukcję ogona rozkładu kosztu (np. p95/p99 kosztu/WU), nie tylko średniej. Jak: symulacje scenariuszowe kalibrowane do danych billingowych + quasi-eksperymenty (np. interrupted time series) po wdrożeniu progów. Dlaczego krytyczne: kaskady są z definicji "rzadkie i duże", więc średnie często kłamią. [5]

## Ocena naukowa planu testów w czterech wymiarach ważności

W "najostrzejszej" klasyfikacji metod empirycznych test musi być oceniony nie tylko "czy działa", ale w czterech kryteriach znanych z tradycji projektowania eksperymentów i quasi-eksperymentów: **statistical conclusion validity**, **internal validity**, **construct validity**, **external validity** (ramy te są klasycznie rozwijane w podręcznikach Shadish–Cook–Campbell). [6]

Oceniam cztery rodziny testów z planu: jednostkowe, integracyjne, obciążeniowe, ekonomiczne.

Testy jednostkowe
Wartość naukowa: bardzo wysoka internal validity i wysoka powtarzalność, bo testują lokalne inwarianty (np. deterministyczność delty, integralność SHA256). Największe ryzyko: niska external validity (nie dowodzą jeszcze amortyzacji kolapsu, tylko poprawność mechanizmu). To nie wada, tylko właściwa rola. Problemem staje się dopiero wtedy, gdy jednostkowe testy są mylone z dowodem systemowym.

Testy integracyjne (E2E)
Wartość naukowa: dobra construct validity, bo obejmują sprzężenia i integracje (tam rodzą się kaskady). Największe ryzyko internal validity: konfuzja środowiskowa (inne wersje zależności, zmienne sieciowe, cache, różnice konfiguracji) i artefakty testów (np. testuje się przypadek, który nie przypomina agentowego "workloadu"). Aby utrzymać rygor, E2E musi mieć kontrolę wersji (SBOM/scan) i deterministyczną konfigurację infrastruktury. To wprost pasuje do tego, co budujesz w `sbom`: event store, snapshot, delta, gate.

Testy obciążeniowe
Wartość naukowa: kluczowa dla Twojej tezy, bo dotyka zjawisk progowych i zmiany reżimu, o których pisze m.in. entity["people","Duncan J. Watts","network scientist"] (rzadkie, duże kaskady wyzwalane małym impulsem w sieciach progowych). [7]
Największe ryzyko: "niewłaściwa dystrybucja obciążenia" (np. zbyt gładka, podczas gdy rzeczywisty agentowy ruch jest bursty), oraz druga pułapka: testy przeciążeniowe mogą być zbyt krótkie, by ujawnić dryf i efekty kumulacji (sustained, pamięć, retry storms). Minimalna poprawka naukowa: modelowanie obciążenia jako mieszaniny (burst + sustained + fault injection), a metryki raportować jako rozkłady

(p50/p95/p99), nie tylko średnie, zgodnie z ostrzeżeniami ASA, że pojedynczy próg statystyczny lub pojedyncza liczba łatwo prowadzi do błędnej interpretacji. [8]

Testy ekonomiczne i scenariuszowe
Wartość naukowa: warunkowa. Monte Carlo i symulacje nie są "dowodem świata", tylko "dowodem konsekwencji założeń", co jest naukowo uczciwe, jeśli (a) parametry są jawne, (b) jest kalibracja do danych, (c) są testy wrażliwości. Ryzyko: bez kalibracji łatwo wpaść w to, co metanauka opisuje jako nadmiar stopni swobody i podatność na fałszywe wnioski (zwłaszcza przy wielu metrykach). [9]
W praktyce "kontrapunkt do rzeczywistości" jest tu szczególnie twardy: jeśli IEA pokazuje, że to energia i bottlenecki infrastrukturalne są realnym constraintem oraz że niepewności scenariuszowe są duże, to test ekonomiczny musi mieć scenariusze (Base/Headwinds/High efficiency) lub przynajmniej szerokie przedziały na koszty i popyt WU. [1]

## Ocena metryk i ryzyko metryczne w warunkach złożoności

W absolutnym reżimie naukowym metryki są częścią hipotezy. Jeśli metryka jest zła, test obala nie hipotezę, tylko pomiar. Dlatego ocena jakości metryk jest równie ważna jak ocena testów. [10]

Metryki techniczne
Definicja operacyjna: latencja p50/p95/p99, throughput, error rates (5xx/429/timeouts), retry rate, ejection rate (outlier detection), saturation CPU/RAM/conn, backlog. Pomiar: telemetryka rozproszona i korelacja sygnałów (logs+traces+metrics) z id żądania / id jednostki pracy. To podejście jest spójne z kierunkiem OpenTelemetry: standard dopinania kontekstu (trace/span id, baggage) do logów i metryk, aby można było jednoznacznie skorelować sygnały w systemach rozproszonych. [11]
Pułapka: średnia latencja bywa myląca przy heavy tails (kaskady), a "przeciętny throughput" nie opisuje przejść fazowych. W Twojej domenie p99 jest często metryką właściwszą niż mean.

Metryki operacyjne
Definicja: MTTR, czas powrotu do reżimu stabilnego, "kaskadowość" (odsetek incydentów propagujących downstream), stabilność bramek (false positive/negative). Pułapki: (a) Goodhart – gdy metrykę zrobisz celem, system zaczyna być "optymalizowany pod wynik", a nie pod rzeczywistość; (b) selection bias – mierzysz tylko to, co dotarło do logów. W literaturze Goodharta i Strathern (w popularnej formule "kiedy miara staje się celem...") główna teza brzmi: regularność statystyczna psuje się pod presją sterowania, więc metryki muszą być projektowane tak, by minimalizować bodźce do "gry". [12]

Metryki ekonomiczne
Definicja: cost per work unit, rozkład kosztu (p50/p95/p99), EV(ryzyka błędu), cashflow impact. Pomiar: eksport billingowy + metering WU + przypisanie kosztów operacyjnych (obsługa incydentów, review, regresje). Pułapki: (a) "cost attribution gap" – jeśli nie ma spójnego id WU w telemetryce, koszt rozlewa się po usługach jak mgła, (b) "multiple comparisons" – wiele KPI naraz daje pozorne "sukcesy" przez przypadek, co ASA opisuje jako typowy błąd interpretacji, gdy p-value i progi są traktowane jako wyrocznia. [13]

Metryki uzupełniające, które podnoszą rygor
Aby metryki były odporne na kaskady i gaming, rekomenduję (jako standard badawczy):
- raportowanie rozkładów i przedziałów (bootstrap/percentyle) zamiast samych średnich, bo kaskady są rzadkie i duże. [14]
- "safety-metric" dla progów: odsetek ruchu odrzuconego kontrolowanie (429) vs odsetek awarii kaskadowych (5xx), aby nie wpaść w iluzję, że "mniej błędów" = "lepiej" (czasem lepiej jest fail-closed). To jest zgodne z dokumentacją Envoy dla `failure_mode_deny` (fail closed) i mechaniką rate limiting.

[15]
- rejestrowanie driftu protokołu (zmienność zachowania "bytu") – Twoje `writeups` trafnie wskazuje dryf jako warunek falsyfikacji predykcji. To jest naukowo ważne, bo eliminuje fałszywe "trafienia" na historii. [16]

## Procedury falsyfikacji, statystyka i wymagane rozmiary próbek

W reżimie Poppera teoria jest naukowa, jeśli istnieją obserwacje, które ją obalą; metodologicznie jednak potrzebujesz procedury pomiaru, która minimalizuje błąd i bias, inaczej "falsyfikacja" jest pozorna. [2]

Warunki falsyfikacji dla rdzenia Twojej tezy
Hipoteza "amortyzujemy kaskady" jest obalona, jeśli po włączeniu progów i bramek spełniony jest którykolwiek z warunków:
- kaskadowość (propagacja błędów) nie maleje, a p99 latencji/awarii rośnie lub MTTR się wydłuża,
- koszt/jednostka pracy (p95/p99) rośnie po wdrożeniu progów (czyli progi generują narzut większy niż redukcja incydentów),
- spada replikowalność kontekstu (zwiększa się "niewyjaśniona wariancja"),
- redukcja błędów jest pozorna, bo system fail-open przepuszcza przeciążenie (np. brak `failure_mode_deny` przy niedostępnej usłudze limitującej, co w Envoy jest jawnie opisane jako ryzyko). [15]

Dobór testów statystycznych
- Dla różnic w proporcjach (np. odsetek kaskadowych awarii): test dwóch proporcji lub model binarny (logit) + przedziały ufności. Praktyczne narzędzia do power/sample size dla dwóch proporcji są dobrze opisane w dokumentacji statystycznej (np. formuły i parametry: α, moc, p1/p2). [17]
- Dla rozkładów heavy-tail (latencje, koszty): testy oparte o percentyle/bootstrapping lub modele odporne (quantile regression), bo mean-test bywa niestabilny w obliczu rzadkich, ekstremalnych zdarzeń, które są istotą kaskad. [14]
- Dla efektu wdrożenia w czasie: interrupted time series (ITS) lub difference-in-differences (DiD) zamiast naiwnego before/after. ITS ma literaturę ostrzegającą, że przy małej liczbie punktów czasowych spada wiarygodność i rośnie ryzyko błędów I rodzaju; zalecenia projektowe obejmują minimalną liczbę punktów oraz ostrożność w interpretacji "borderline significance". [18]
- Dla DiD kluczowe jest założenie parallel trends / common shocks i testy pre-trendów jako fakt empiryczny, nie jako gwarancja; istnieją opracowania, które explicytnie rozbijają to założenie i ostrzegają przed fałszywym poczuciem bezpieczeństwa. [19]

Szacunki rozmiaru próby
Ponieważ nie mamy tu jeszcze danych produkcyjnych, podaję szacunki jako "rzędy wielkości" (zależne od baseline i oczekiwanego efektu), a nie jako liczby absolutne:

- Jeśli wskaźnik kaskadowych awarii ma baseline rzędu 5–10% i chcesz wykazać spadek o kilka punktów procentowych, typowe kalkulacje mocy dla dwóch proporcji wskazują, że mogą być potrzebne setki do tysięcy obserwacji na wariant (zależnie od wielkości efektu i α). [20]
- Dla ITS: literatura symulacyjna wskazuje, że przy krótkich szeregach czasowych rośnie błąd; zalecenia projektowe mówią o minimum kilkunastu–kilkudziesięciu punktach czasowych (np. rekomendacje "minimum 24 punkty" w pewnych analizach ITS) oraz o zależności mocy od efektu i autokorelacji. [18]
- Dla metryk latencji/obciążenia: ponieważ dystrybucje bywają ciężkoogonowe, bardziej sensowne jest planowanie próby tak, aby stabilnie estymować p95/p99 (co zwykle wymaga większej liczby requestów niż do oszacowania średniej).

Kontrola wielokrotnych porównań
W planie masz wiele metryk. Jeżeli testujesz wiele hipotez, musisz kontrolować familywise error lub FDR. ASA wprost wskazuje, że wnioskowanie wymaga pełnego raportowania i transparentności zamiast "magii progu p-value". [21]
Jednocześnie metody FDR mają własne ryzyka interpretacyjne (np. porównywanie różnych eksperymentów), co opisuje literatura metodyczna. [22]

## Plan quasi-eksperymentów, bezpieczeństwo wdrożeń i szablony raportów

W Twojej domenie pełna randomizacja bywa trudna, ale rygor jest osiągalny przez połączenie: feature flags, mirror traffic, stop-conditions oraz projektów quasi-eksperymentalnych (ITS/DiD), o których wspomina literatura projektowania przyczynowego. [23]

Plan wdrożenia A/B i quasi-eksperymentów
- Feature flags dla progów runtime (rate limit, circuit breaker, retries/timeouts). Stop condition: automatyczny rollback, gdy przekroczony jest error budget albo p99 latencji rośnie powyżej progu.
- Mirror traffic: nowa konfiguracja dostaje kopię ruchu, ale nie wpływa na użytkowników (wysoka internal validity dla obserwowalności i zachowania pod obciążeniem).
- ITS dla kosztu/WU i EV(ryzyka): włącz/wyłącz bramkę (gate) w kontrolowanych oknach czasu, zbierając serię punktów przed i po; interpretacja zgodnie z zaleceniami ITS (autokorelacja, liczba punktów). [18]
- DiD dla klientów/modułów: jeśli masz "podobne" strumienie pracy (np. moduły o podobnym profilu), stosujesz progi w jednej grupie, druga pozostaje kontrolą; kluczowa weryfikacja: parallel pre-trends. [24]

Szablony raportów wynikowych
Szablon runu (unit/integration/load) powinien raportować jednocześnie trzy warstwy:
- techniczne: p50/p95/p99, error classes, retry, outlier ejections, saturation,
- operacyjne: MTTR, kaskadowość, false gate rate,
- ekonomiczne: cost/WU i jego percentyle, EV(ryzyka) i jego komponenty.
Zgodnie z ASA raport powinien zawierać nie tylko "czy istotne", ale wielkość efektu, niepewność i jawny opis analizy. [21]

Tabela porównawcza modelu SaaS vs anty-wzorzec
Poniższa tabela ma status "hipotezy konstruktu": pokazuje, co testy mają potwierdzić/obalić.

| Wymiar | Klasyczny SaaS w reżimie AI+Cloud | Anty-wzorzec z Twoich repo |
|---|---|---|
| Metering | Licznik seat/plan; koszt agentowy bywa ukryty | Jednostka pracy (WU) ma być policzalna i przypisywalna kontekstowo |
| Gating | Brak twardych progów; reakcja post-hoc (on-call) | Gate w CI (SBOM/scan) + progi runtime (rate limit / circuit breaker) |
| Kontekst | Słaba replikowalność; logi bez spójnego kontekstu | Protokół kontekstu (UUID+hash) + korelacja telemetryki (trace context) |
| Przenaszalność | Integracje rosną organicznie | Mapowanie woluminu QV9D→artefakty, aby dało się audytować i sterować |

| Wymiar | Klasyczny SaaS w reżimie AI+Cloud | Anty-wzorzec z Twoich repo |
|---|---|---|
| Koszt | Średnie KPI maskują ogon ryzyka | Cel: obcięcie ogona (p95/p99) i redukcja EV(straty) |

## Ograniczenia, ryzyka i końcowe wnioski falsyfikowalne

Ograniczenia badania
Najważniejsze ograniczenie jest epistemiczne, nie techniczne: repozytoria zawierają mieszankę artefaktów implementacyjnych (pipeline, polityki) i specyfikacyjnych (protokół 9D, mapowanie), a to oznacza, że dowód "działa ekonomicznie" wymaga danych kosztowych i obserwacji długookresowej. To nie unieważnia tezy, ale klasyfikuje obecny etap jako "program badawczy + implementacja mechanizmów sterowania", a nie "empirycznie udowodniona amortyzacja gospodarcza".

Ryzyka metodologiczne
- Ryzyko Goodharta: jeśli KPI becomes target, system będzie "upiększany" pod metrykę; dlatego metryki muszą być odporne na gaming i raportowane jako rozkłady. [12]
- Ryzyko fałszywych wniosków przy wielu testach i elastycznej analizie; metanauka ostrzega, że bez dyscypliny (prerejestracja, ograniczenie stopni swobody) rośnie udział false positives. [9]
- Ryzyko external validity: nawet najlepsze sterowanie lokalne nie "unieważnia" constraintów energii i lead time'ów infrastruktury; IEA pokazuje, że to realny, scenariuszowy czynnik. [1]

Wnioski falsyfikowalne po wykonaniu testów
Po wykonaniu zaproponowanych eksperymentów będziesz w stanie wydać wnioski o statusie naukowym (w sensie Poppera i ważności przyczynowej), w formie:

- H1 (progi amortyzują kaskady) przyjęta/obalona na podstawie zmiany kaskadowości, MTTR i reżimu degradacji (429 vs 5xx) przy kontrolowanych workloadach. Wsparcie teoretyczne: kaskady progowe w sieciach i rzadkie global cascades. [7]
- H2 (kontekst jest dowodowy) przyjęta/obalona na podstawie integralności (hash), replikowalności (replay) i zdolności wyjaśniania różnic przez delta-kontekstu, zgodnie z rygorem "basic statements" i praktyką metodologiczną falsyfikacji. [2]
- H3 (metering WU stabilizuje ekonomię) przyjęta/obalona na podstawie rozkładów cost/WU i EV(straty) w ITS/DiD oraz na podstawie odporności w scenariuszach energii (kalibrowanych do realnych danych billingowych). [25]

Praktyczne rekomendacje "najwyższej dźwigni"
- Prerejestruj primary endpoints (np. kaskadowość, MTTR, p99 latencji, p95/p99 cost/WU) i analizę; bez tego łatwo o p-hacking i "zwycięstwo narracyjne". [26]
- Wprowadź standard korelacji telemetryki WU (trace context) w stylu OTel, bo bez tego nie domkniesz meteringu ekonomicznego. [11]
- Kalibruj progi na podstawie kosztu/WU (nie tylko RPS), bo w AI to WU jest nośnikiem kosztu. To jest zgodne z kierunkami FinOps rozszerzającymi zakres na SaaS i data center oraz z realnym constraintem energii rosnącym w scenariuszach IEA. [27]
- Dla ITS zapewnij wystarczającą liczbę punktów czasowych i raportuj autokorelację; literatura wskazuje, że krótkie szeregi podnoszą ryzyko błędnych wniosków. [18]

---

[1] https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai
https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

2 https://plato.stanford.edu/entries/popper
https://plato.stanford.edu/entries/popper

3 https://www.youtube.com/watch?v=VDrcgEne6lU
https://www.youtube.com/watch?v=VDrcgEne6lU

4 5 7 14 https://pmc.ncbi.nlm.nih.gov/articles/PMC122850/
https://pmc.ncbi.nlm.nih.gov/articles/PMC122850/

6 23 https://lawcat.berkeley.edu/record/365766
https://lawcat.berkeley.edu/record/365766

8 10 13 21 26 https://www.amstat.org/asa/files/pdfs/p-valuestatement.pdf
https://www.amstat.org/asa/files/pdfs/p-valuestatement.pdf

9 16 https://colab.ws/articles/10.1371%2Fjournal.pmed.0020124
https://colab.ws/articles/10.1371%2Fjournal.pmed.0020124

11 https://opentelemetry.io/docs/reference/specification/logs/
https://opentelemetry.io/docs/reference/specification/logs/

12 https://www.damtp.cam.ac.uk/user/mem/papers/LHCE/goodhart.html
https://www.damtp.cam.ac.uk/user/mem/papers/LHCE/goodhart.html

15 https://www.envoyproxy.io/docs/envoy/latest/configuration/other_protocols/thrift_filters/rate_limit_filter
https://www.envoyproxy.io/docs/envoy/latest/configuration/other_protocols/thrift_filters/rate_limit_filter

17 power twoproportions — Power analysis for a two-sample proportions test
https://www.stata.com/manuals/pss-2powertwoproportions.pdf?utm_source=chatgpt.com

18 25 Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study | BMC Medical Research Methodology | Full Text
https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-021-01364-0?utm_source=chatgpt.com

19 Difference-in-Differences
https://diff.healthpolicydatascience.org/?utm_source=chatgpt.com

20 manual:two_sample_proportion_equal_sample_size [WebPower WIKI]
https://webpower.psychstat.org/wiki/manual/two_sample_proportion_equal_sample_size?utm_source=chatgpt.com

22 note on the false discovery rate and inconsistent comparisons between experiments | Bioinformatics | Oxford Academic
https://academic.oup.com/bioinformatics/article-abstract/24/10/1225/178402?utm_source=chatgpt.com

24 Introduction to Difference-in-differences Design
https://www.jstage.jst.go.jp/article/ace/3/3/3_74/_html/-char/en?utm_source=chatgpt.com

27 https://www.finops.org/insights/2025-finops-framework/
https://www.finops.org/insights/2025-finops-framework/