GDP.pdf Review Report

Author: Benchmarks' Last Exam Lab 2026-10-15

GDP.pdf tests how well models can read and reason about PDF documents from ten professional fields. Each question asks the model to find and use evidence in a document to solve a problem from work in areas such as finance, engineering or research.

Abstract

We reviewed the 100 publicly available tasks: **40 Accepted**, **13 Doubtful** and **47 Refused**.

Most confirmed issues concern reference answers that conflict with the source or grading requirements absent from the prompt. Some criteria turn a rule against wrong answers into a ban on accurate context or explanations of excluded options, causing correct work to lose credit.

Review outcome	Tasks	Basis for assessment
Accepted	**40**	Within the scope reviewed, no material defect or concrete unresolved concern was found.
Doubtful	**13**	A specific concern remains unresolved: its interpretation, trigger or effect on scoring needs verification.
Refused	**47**	At least one defect affecting evaluation validity was confirmed and survived independent counter-review.

Scope and version

The report covers 100 tasks in the openly available copy. Fixed dataset version: 7f76578

The scoring implementation uses a fixed version of surge-ai/gdp-pdf: 7a72a51

Report updated on November 4, 2026.

Distribution of findings

Issue type	Refused	Doubtful	Total
C1 Incorrect answers or scoring logic	21	4	25
C2 Scoring requirements not stated in the task	25	9	34

Issue type	Refused	Doubtful	Total
C3 Missed critical errors	1	0	1
C4 Scoring bypasses or answer leakage	0	0	0
C5 Other issues	0	0	0
Total	**47**	**13**	**60**

Each task is counted once, under its primary issue category.

Representative cases

A one-percentage-point decrease becomes 1,000 basis points

C1 Incorrect answers or scoring logic

Task: gdp-060

The task asks for an analysis of Airbnb's financial report, including changes in cost of revenue and administrative expenses as shares of revenue. Those changes must be expressed in basis points. One basis point is 0.01 percentage points.

The table on PDF page 71 shows cost of revenue falling from 17% of revenue in 2023 to 16% in 2024. Using these displayed percentages, the decrease is one percentage point, or 100 basis points. Criterion r9 instead requires a decrease of 1,000 basis points.

A decrease of 1,000 basis points is ten percentage points, which does not match this cost line. The finding is an incorrect reference value.

Evidence 2

- Cost table on PDF page 71
- Prompt and r9 in dataset row 60

The public prompt never states the analysis task

C2 Scoring requirements not stated in the task

Task: gdp-009

The public prompt is an editing note: an issue can be resolved by changing the prompt, context or rubric, with details elsewhere. It never asks the solver to perform an analysis. The attachment contains ten photographs and a map, with no additional question.

Criterion r1 nevertheless requires three judgments for every photograph: whether it shows hail damage, whether that damage can be attributed to the May 8, 2017 storm, and whether poor maintenance is evident.

These requirements are not available to the solver. Recognizing what is in the photographs does not reveal that these three questions must be answered for each image. The core task definition is missing from the public input.

Evidence 2

- Photographs and map
- Public prompt and r1 in dataset row 9

Correctly explaining excluded age groups loses credit

C2 Scoring requirements not stated in the task

Task: gdp-064

The task asks for child age groups whose anemia prevalence is at least 70% in both urban and rural areas, followed by each group's combined prevalence. The qualifying groups are 6–11 months and 12–23 months, with combined rates of 88.1% and 73.7%.

The normal answer compared five age groups, explicitly marked the other three as not qualifying, and reported final results only for the two eligible groups. Criterion r5, however, turns a restriction on selecting other groups into a restriction on mentioning them.

The grader rejected this accurate exclusion explanation. The other seven criteria passed, giving a mean score of 0.875 but no all-pass result. The error concerns an added restriction on how to explain the answer, not an incorrect selection of age groups.

Evidence 3

- Combined rates on PDF page 8
- Urban/rural chart on page 10
- Prompt and r5 in dataset row 64

The entire deficiency-effects column can be wrong

C3 Missed critical errors

Task: gdp-094

The task asks for a two-column table of nutrients needed during pregnancy and the effects of deficiency, drawn from the attached document for pediatric nursing colleagues. The source first discusses iron, folate and iodine deficiencies during pregnancy. A separate sentence then discusses childhood vitamin A deficiency and blindness.

The relevant criteria check the table, column headings, the three nutrient names and the absence of additional supplement recommendations. They do not check the effects column. Replacing all three pregnancy-related effects with blindness from the adjacent childhood example leaves those written conditions satisfied; the other criteria do not examine this change either.

The missing check concerns the nutrient–effect relationships explicitly requested by the task.

Evidence 2

- Nutrition box on PDF page 10
- Prompt and all 21 criteria in dataset row 94

Task-by-task findings

Refused: 47 tasks

gdp-002

C2 Scoring requirements not stated in the task

The prompt asks only for the pages containing fee information. The rubric additionally requires all four fee categories to be restated. Correct page references with genuine examples can still fail this extra requirement.

gdp-004

C2 Scoring requirements not stated in the task

The transformer and part number 60536 were correct. An accurate additional reference to a 120V diagram on another page nevertheless caused the only failed check, leaving a mean score of 0.75.

gdp-006

C1 Incorrect answers or scoring logic

The reference assigns a PESI risk score of 138 to Class IV. Both the attached table and the original study place it in Class V. A correct classification conflicts with the grading target.

gdp-009

C2 Scoring requirements not stated in the task

The public prompt contains only an editing note, and the attachment supplies no question. The rubric nevertheless requires a hail-damage, storm-attribution and maintenance analysis for each of ten photographs. The core task was never communicated to the solver.

gdp-013

C2 Scoring requirements not stated in the task

The answer placed people with unknown fitness in a separate Unknown category rather than mixing them into the fitness gradient. Their presence in the table still caused its only failed check, although the prompt did not prohibit this accurate clarification.

gdp-015

C1 Incorrect answers or scoring logic

The rubric requires the insurer to have a duty to defend, overlooking the attached exclusion for assault- and battery-related suits, including negligence theories. A conditional analysis that follows the policy conflicts with this fixed conclusion.

gdp-018

C1 Incorrect answers or scoring logic

The rubric combines water immersion and accelerated aging into one condition. The source gives hardness changes of ±10 for immersion and +15/−0 for aging. This mismatch caused credit to be withheld for a correctly answered immersion requirement.

gdp-021

C2 Scoring requirements not stated in the task

The prompt permits ranking suppliers by dates and/or problems. The rubric forbids ranking by dates alone and fixes the worst supplier, excluding a different date-based ranking supported by the source table.

gdp-022

C1 Incorrect answers or scoring logic

The rubric reverses the load-bearing relationship, requiring recognition that an extractor hood can support a wall. The source instead makes the hood an exception to the restriction on wall cabinets on a weak wall; it does not support the reversed claim.

gdp-023

C1 Incorrect answers or scoring logic

The task asks which project category has the largest increase. The reference accepts Power & Regional, but the chart shows a larger increase for Neighborhood & Community between the end of 2023 and the end of 2025.

gdp-024

C2 Scoring requirements not stated in the task

The rubric treats eight months at the current agency as only eight months of qualifying federal service in total. The prompt does not rule out previous service, and the handbook allows the required 12 months to accumulate across agencies.

gdp-025

C1 Incorrect answers or scoring logic

The rubric requires reporting that the drawings lack a roof pitch. Page 2 of the attachment already labels both roof slopes 6:12, so a correct drawing review conflicts with the expected missing-item claim.

gdp-026

C2 Scoring requirements not stated in the task

The report does not explain how pending registrations enter the valuation model, yet the rubric requires predicting that the valuation will stay unchanged. A reasoned answer that the direction cannot be determined fails this fixed conclusion.

gdp-028

C2 Scoring requirements not stated in the task

The normal answer quoted a real heading containing Mr. Henley in an accurate page reference. It did not identify him as an executive vice president, but the mere appearance of his name still caused a failed check.

gdp-031

C2 Scoring requirements not stated in the task

The rubric silently restricts HFCS to two finished products with that name, excluding an interpretation based on the high-fructose corn syrup content of recipes. The two readings change production-line eligibility for 28 products, a core scheduling constraint.

gdp-036

C2 Scoring requirements not stated in the task

A complete, correct four-item list can violate the rubric merely by adding an accurate drawing-version reference. The prompt does not ban such source information. Neither recorded run triggered this rejection path.

gdp-040

C1 Incorrect answers or scoring logic

The reference part number for a ceramic bearing's outer ring is 316L616. The manufacturer's main drawing and both outer-ring symbols show 361L616. Copying the correct number conflicts with three rubric requirements.

gdp-041

C2 Scoring requirements not stated in the task

The normal answer distinguished tenure before and after childbirth and supplied the correct medical costs. It still lost credit for including the 0–4-year tenure bracket, which the public instructions did not prohibit.

gdp-042

C1 Incorrect answers or scoring logic

The rubric turns a slower rate of BMI increase in the 18–20 age group into no increase or change in BMI. The source table shows that BMI was still rising, just more slowly, confusing a change in level with a change in growth rate.

gdp-043

C2 Scoring requirements not stated in the task

The rubric requires identifying and moving or avoiding a recessed ceiling speaker that the public materials do not establish. An answer faithfully identifying the equipment actually shown can still miss this required fact.

gdp-045

C2 Scoring requirements not stated in the task

The word type can reasonably refer to the PDF's explicit TYPE column, yielding 13 first-floor double-hung windows. The rubric instead silently interprets it as a MARK number and accepts only zero.

gdp-048

C2 Scoring requirements not stated in the task

An open preliminary electrical-planning request is restricted by the rubric to 12–16 circuits and at least 12–14 board ways. It excludes a reasoned alternative with eight current circuits and two spare ways.

gdp-052

C2 Scoring requirements not stated in the task

The rubric requires advice about a statutory demand letter at least 30 days before filing suit. That procedural rule is absent from the attachment, while the prompt prohibits outside legal material.

gdp-053

C2 Scoring requirements not stated in the task

The rubric unconditionally bans inverse-square-law calculations. This excludes a reasonable adjustment from the study's 50-foot far-field data to 15 metres, even when the estimate meets the numerical requirement. The recorded runs did not test this rejection path.

gdp-056

C1 Incorrect answers or scoring logic

The rubric treats Kokkonen as the only shared citation, rejecting an accurately identified lower-court decision, 103 F.4th 294, that both documents also cite. The normal answer lost credit for this addition and failed the all-pass result.

gdp-057

C1 Incorrect answers or scoring logic

The rubric restricts the diseases associated with four intermediate-frequency models to PTSD, omitting four other diseases in Table 5. It consequently requires two drugs that are not shared by all the associated diseases.

gdp-060

C1 Incorrect answers or scoring logic

The financial report's displayed cost-to-revenue ratio falls from 17% to 16%, a decrease of 100 basis points. The reference requires 1,000 basis points, overstating this line item's change by a factor of ten.

gdp-061

C2 Scoring requirements not stated in the task

The normal answer supplied a complete field list but added ordinary grouping headings for readability. An undisclosed ban on these headings caused a failed check, making formatting an extra condition for accepting correct content.

gdp-062

C1 Incorrect answers or scoring logic

The rubric treats Chloe's absence from the named-insured list as sufficient to rule out all basic first-party medical coverage. It overlooks the policy's conformity-to-law clause and relevant resident-relative coverage rules.

gdp-064

C2 Scoring requirements not stated in the task

The answer selected the correct two age groups and accurately explained why three others did not qualify. The rubric expanded 'do not select other groups' into 'do not mention them,' causing the only failed check and a mean score of 0.875.

gdp-067

C1 Incorrect answers or scoring logic

The reference dates the air-conditioner A-coil replacement to 2020. Page 4 of the PDF explicitly records September 2022, so correctly extracting the date conflicts with the grading target.

gdp-070

C1 Incorrect answers or scoring logic

The reference accepts a 2019 mean sale price of $236,227.50, instead of the correct $229,241.67. The erroneous value is reproduced by including a 2018 sale in the 2019 group.

gdp-071

C1 Incorrect answers or scoring logic

The rubric requires describing the downslope stockpile silt fence as unclear or only partially provided. Drawing C-3 already shows and labels it SF, so correctly identifying the existing fence conflicts with the reference assessment.

gdp-073

C1 Incorrect answers or scoring logic

The rubric merges five four-base CTGC repeats into the three-base TGC group, requiring counts of CTGC=0 and TGC=12. Reading the underlined source sequences correctly gives five and seven respectively.

gdp-076

C2 Scoring requirements not stated in the task

The rubric prohibits a percentage increase in male relative mortality, excluding a normal comparison supported by the chart and explicitly labelled approximate. The prompt does not impose this ban, and the recorded runs did not trigger it.

gdp-079

C1 Incorrect answers or scoring logic

The task asks about the ECG in acute ST-elevation myocardial infarction, but the rubric requires depression in several leads. It excludes the corresponding elevation pattern supported by primary cases and cohort evidence. The normal run did not demonstrate rejection on these four checks.

gdp-080

C1 Incorrect answers or scoring logic

The rubric applies a leave accrual rate of 1.75 days per month, available only after 15 years of service, to the entire preceding two years. It requires 42 days, whereas applying the rates to the stated timeline normally gives about 36.

gdp-082

C1 Incorrect answers or scoring logic

The rubric requires saying that the paper gives no explicit methodological advantages or disadvantages. The first-page abstract states advantages of deep learning and both strengths and limitations of the two-layer analytical fit.

gdp-087

C2 Scoring requirements not stated in the task

The prompt asks for weak recommendations from all tables. The rubric silently limits the scope to Tables 2A–2D and bans three items in other tables that also carry the Category II label. Complete extraction therefore conflicts with grading.

gdp-089

C2 Scoring requirements not stated in the task

Claim 5 omits the first actual-cash-value payment date and the date repair costs were incurred, so the policy's 180-day period cannot be calculated. The rubric nevertheless fixes a denial of the hold-back payment, substituting two different dates.

gdp-091

C1 Incorrect answers or scoring logic

For a second-quarter profitability comparison, the reference uses the industry's first-half share of unprofitable institutions, 6.30%. The same-quarter figure is 6.59%, so the gap from community banks' 6.75% is overstated.

gdp-093

C2 Scoring requirements not stated in the task

Table 16 lists improvement opportunities, not a ranking of current standards' strengths. The rubric nevertheless requires naming Mfg Pyramid as currently strongest, treating opportunity coverage as existing capability.

gdp-094

C3 Missed critical errors

The task asks for the consequences of nutrient deficiencies during pregnancy, but the rubric checks table structure and nutrient names without checking the consequences. Replacing the whole column with blindness from an adjacent childhood example still satisfies the relevant written conditions. This answer was not run through the judge.

gdp-097

C2 Scoring requirements not stated in the task

The requested template calls for hourly production forecasts. The normal answer provided independently reproducible forecasts of 1,595.8 and 1,665.5 boxes per hour, but grading also required weekly totals and rejected these two items.

gdp-098

C1 Incorrect answers or scoring logic

The rubric treats records involving public funds as categorically ineligible for the attorney-communication exemption. It omits the source's non-privileged-content condition and the protection retained for confidential legal billing material.

gdp-099

C2 Scoring requirements not stated in the task

The rubric limits machining-parameter answers to general formulas, excluding qualitative speed and feed advice permitted by the PDF. The normal answer gave no specific rate values but still lost credit for such advice.

gdp-100

C2 Scoring requirements not stated in the task

The normal answer used and explained the paper's own R/I/E mechanism names. A rubric condition banned these terms and rejected the answer, although the public task did not prohibit source-supported terminology.

Doubtful: 13 tasks

gdp-019

C2 Scoring requirements not stated in the task

A computed value of 19.275 rounds to 19.28 at two decimal places, while the reference says 19.27. Likewise, 94.53% may exceed a literal upper bound of 94.5%. The recorded answers did not correctly submit these values, so the acceptance range remains unconfirmed.

gdp-029

C2 Scoring requirements not stated in the task

The rubric may prohibit supporting monetary data from another chart in the paragraph about headcount, even when the support is clearly labelled and the headcount analysis is correct. Whether the public task requires this strict separation remains unresolved.

gdp-030

C2 Scoring requirements not stated in the task

The instrument's error bound is ±50 ppm, while some CO_2 readings are barely above 50 ppm. The rubric fixes a conclusion about six-hour pH control, but different underlying conditions fit that error bound, so the evidence may not determine a unique conclusion.

gdp-033

C2 Scoring requirements not stated in the task

The prompt omits prior periodontal-treatment history, but the rubric fixes a 70% patient share for one service. A qualified answer received credit, weakening the false-rejection concern; whether the percentage depends on missing facts remains unresolved.

gdp-047

C2 Scoring requirements not stated in the task

The prompt asks for the numerical difference without specifying an absolute difference or subtraction order. Both 17/22 and −17/−22 have plausible interpretations, but the rubric fixes the negative values and expressly bans the positive ones.

gdp-058

C2 Scoring requirements not stated in the task

Compensation changes around an executive transition can be calculated by office or by individual. The normal answer explicitly used year-end officeholders and added same-person comparisons, but lost credit for not using the rubric's entry/exit convention. Its necessity remains unestablished.

gdp-063

C2 Scoring requirements not stated in the task

The answer correctly reported the target years and labelled two additional rows as baseline context. This caused its only failed check. The prompt does not expressly ban accurate background, but whether the extra rows exceed its requested scope remains disputed.

gdp-066

C1 Incorrect answers or scoring logic

The rubric describes cruise operating days' 5% contribution to revenue growth as a 5% increase in operating days themselves. Other correct operating metrics could legitimately satisfy the check, so this confusion's independent effect on credit has not been isolated.

gdp-074

C2 Scoring requirements not stated in the task

The prompt specifies an expanded steam-reformer boundary, while the rubric fixes whether streams 401, 428 and 430 are inputs. The recorded answer contains other mismatches and omissions, so its lost credit cannot yet be attributed to the boundary interpretation.

gdp-083

C1 Incorrect answers or scoring logic

It remains unclear how using an RV as a permanent residence affects uninsured-motorist bodily-injury coverage under this policy. The normal answer's conditional analysis failed three checks, while another accepted the qualification.

gdp-085

C1 Incorrect answers or scoring logic

Stable overall cell expansion and increased EdU proliferation-assay results can coexist, but the rubric does not clearly distinguish them. Two local failures strengthen the concern, though a complete correct answer may still satisfy the condition.

gdp-088

C1 Incorrect answers or scoring logic

The rubric gives competitiveness as an example of a bad explanation, potentially penalizing a source-supported goal of competitiveness in a specific field. The full context permits a compatible reading, and the recorded answers did not test this wording.

gdp-092

C2 Scoring requirements not stated in the task

The public question asks for three promising mitigation techniques, while the rubric fixes a five-item candidate list. The recorded runs did not test whether a reasonable third choice outside that list would be rejected.

Accepted: 40 tasks

gdp-001
gdp-003
gdp-005
gdp-007
gdp-008
gdp-010
gdp-011
gdp-012
gdp-014
gdp-016
gdp-017
gdp-020
gdp-027
gdp-032
gdp-034
gdp-035
gdp-037
gdp-038
gdp-039
gdp-044
gdp-046
gdp-049
gdp-050
gdp-051
gdp-054
gdp-055
gdp-059
gdp-065
gdp-068
gdp-069

gdp-072

gdp-075

gdp-077

gdp-078

gdp-081

gdp-084

gdp-086

gdp-090

gdp-095

gdp-096

Effects on evaluation and follow-up

Document-QA criteria need to follow the source and accept reasonable answers permitted by the prompt. Wrong numbers, dates or entities can reverse the relationship between correct extraction and credit. Hidden word bans and exclusions can turn accurate explanations into reasons for rejection.

The domain-specific judgments still require further verification by experts. Each issue's effect on actual scoring depends on its trigger conditions. We will update the relevant conclusions when verifiable counterevidence, author clarifications or revisions become available.

Contact Us

benchmarks-last-exam.com

Benchmarks-Last-Exam@outlook.com

HarmonyOS Sans SC · Copyright 2021 Huawei Device Co., Ltd.