

Duplicate pages challenge set

Answer key

Results as of October 2026.

This key goes with the 19-page challenge set. For every page it says whether a duplicate detector should flag it for removal or keep it, what PDFHaul's Remove Duplicates tool does with it today, and why.

"Flag" means the page repeats an earlier page and can be removed without losing content. "Keep" means the page is an original, or it only looks like another page and removing it would lose content.

Summary

Should be flagged (8 pages): 2, 3, 6, 9, 10, 13, 14, 18.

Should be kept (11 pages): 1, 4, 5, 7, 8, 11, 12, 15, 16, 17, 19.

PDFHaul today flags 7 of the 8 and none of the 11. The one it misses is page 18, the rescanned page in Section F.

PDFHaul shows two labels. "Duplicate" means the page renders identically to an earlier page. "Likely duplicate" means the text and graphics match but the rendering differs; these are shown for you to review before anything is removed.

Section A: same wording, three PDF generators (pages 1 to 4)

Three pages carry identical wording but were produced by different PDF generators, so the files differ internally and the pages do not render pixel for pixel the same.

Page	Expected	PDFHaul today	Why
1	Keep	Kept	The original. The first copy of a page is always the one that stays.
2	Flag	Likely duplicate of 1	Same wording as page 1, produced by a different generator.
3	Flag	Likely duplicate of 1	Same wording as page 1, produced by a third generator.
4	Keep	Kept	Decoy. Same layout and generator, but the wording is different.

Section B: recompressed image (pages 5 to 7)

All three pages have the same heading and a picture. Two of the pictures are the same image saved at different compression levels.

Page	Expected	PDFHaul today	Why
5	Keep	Kept	The original.
6	Flag	Likely duplicate of 5	Same picture as page 5, recompressed. The content is the same; only the image encoding changed.
7	Keep	Kept	Decoy. Same heading, but a different picture. Text alone cannot tell this page from page 5.

Section C: shifted and rotated (pages 8 to 11)

The same page content appears moved on the page and turned sideways.

Page	Expected	PDFHaul today	Why
8	Keep	Kept	The original.
9	Flag	Likely duplicate of 8	Same content as page 8, shifted on the page.
10	Flag	Likely duplicate of 8	Same content as page 8, on a page turned 90 degrees.
11	Keep	Kept	Decoy. Turned the same way as page 10, but the wording is different.

Section D: blank pages (pages 12 to 14)

Three identical blank pages. They are intentionally empty and carry no section label in the challenge set.

Page	Expected	PDFHaul today	Why
12	Keep	Kept	The first blank page. Remove Duplicates keeps one copy; Remove Empty Pages would remove all three.
13	Flag	Duplicate of 12	Identical to page 12.
14	Flag	Duplicate of 12	Identical to page 12.

Section E: visually similar, substantively different (pages 15 and 16)

Two invoices with the same layout. This is the case where a wrong match loses real data.

Page	Expected	PDFHaul today	Why
15	Keep	Kept	Invoice INV-1042 for \$100.00.
16	Keep	Kept	Invoice INV-1043 for \$100.50. Different invoice number and amount, so it is a different document and must not be flagged.

Section F: simulated rescans (pages 17 to 19)

Image-only pages with no text layer, as produced by a scanner. Two are scans of the same sheet.

Page	Expected	PDFHaul today	Why
17	Keep	Kept	The original scan.
18	Flag	Not detected yet	A second scan of the same sheet as page 17, slightly tilted and with scanner noise. It is a duplicate, but PDFHaul does not match scanned pages unless they are pixel identical.
19	Keep	Kept	Decoy. Same layout as page 17, but one line of text is different.

Try it yourself: upload the challenge set at <https://www.pdfhaul.com/remove-duplicates>