Source: Sci-Hub (14.4%)
- Abstract Writing (2K)
- Paper Titling (2K)
- Caption Writing (2K)
- Experiment Writing (2K)
- Translation (3K)
- MT-QA (interleave) (49.8K)
- MT-QA (image) (49.8K)

Source: Arxiv (55.8%)
- MT-QA (interleave) (211K)
- MT-QA (image) (211K)

Source: OpenReview (29.8%)
- Review (147.6K)
- Reply (77.7K)

Multi-Page Document QA (78.8%)
- Docmatix (141K)
- ArxivQA (31K)
- DUDE (27K)
- MP-DocVQA (51K)
- Doc-750K (750K) (ours)

Single-Page Document QA (10.8%)
- DocVQA (56K)
- DocReason (25.8K)
- InfoVQA (25.5K)
- ChartQA (30.2K)

Multi-Image General QA (3.6%)
- MMDU (45K)

Pure-Text QA (6.8%)
- LongAlign (5.9K)
- LongCite (13.5K)
- LongReward (10K)
- LongAlpaca (12K)
- LongQLoRA (38.7K)

(a) Data Distribution of Doc-750K

(b) Data Recipe for Supervised Fine-tuning