**Purpose:** To outline the sampling strategy for annotating motor finance documents, using custom sample sizes and a 60:40 train/test split, ensuring a representative dataset for fine-tuning and evaluating the text extraction pipeline.

#### **1. Introduction**

Manually labeling all 245,639 motor finance documents in our dataset is impractical due to resource constraints. We select a statistically significant sample to annotate, ensuring the labeled data reflects the entire dataset. This report details the sampling strategy, incorporating custom sample sizes per page count and a 60:40 train/test split. It includes per-page error margins with the margin of error formula in a proper document format, explaining the significance of the 750-document sample for fine-tuning and evaluating our pipeline, which extracts key information (e.g., agreement numbers, customer names) and tables.

#### **2. Document Distribution**

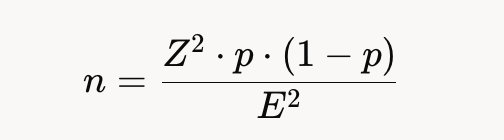
The dataset comprises 245,639 documents, categorised by page count. We focus on documents with 1 to 5 pages, as longer documents (6+ pages) are rare (2,410 documents, 1.8% of total) and less relevant to our current use case:

| Page Count | Total Documents |
| --- | --- |
| 1 | 141,744 |
| 2 | 61,640 |
| 3 | 27,295 |
| 4 | 10,304 |
| 5 | 2,245 |
| Total (1–5 Pages) | 243,228 |

**Decision to Cap at 5 Pages:** Documents with 6+ pages are statistically insignificant (1.8% of total) and likely outliers in motor finance agreements, which are typically 1–5 pages. Focusing on this subset simplifies annotation while covering 98.2% of the dataset.

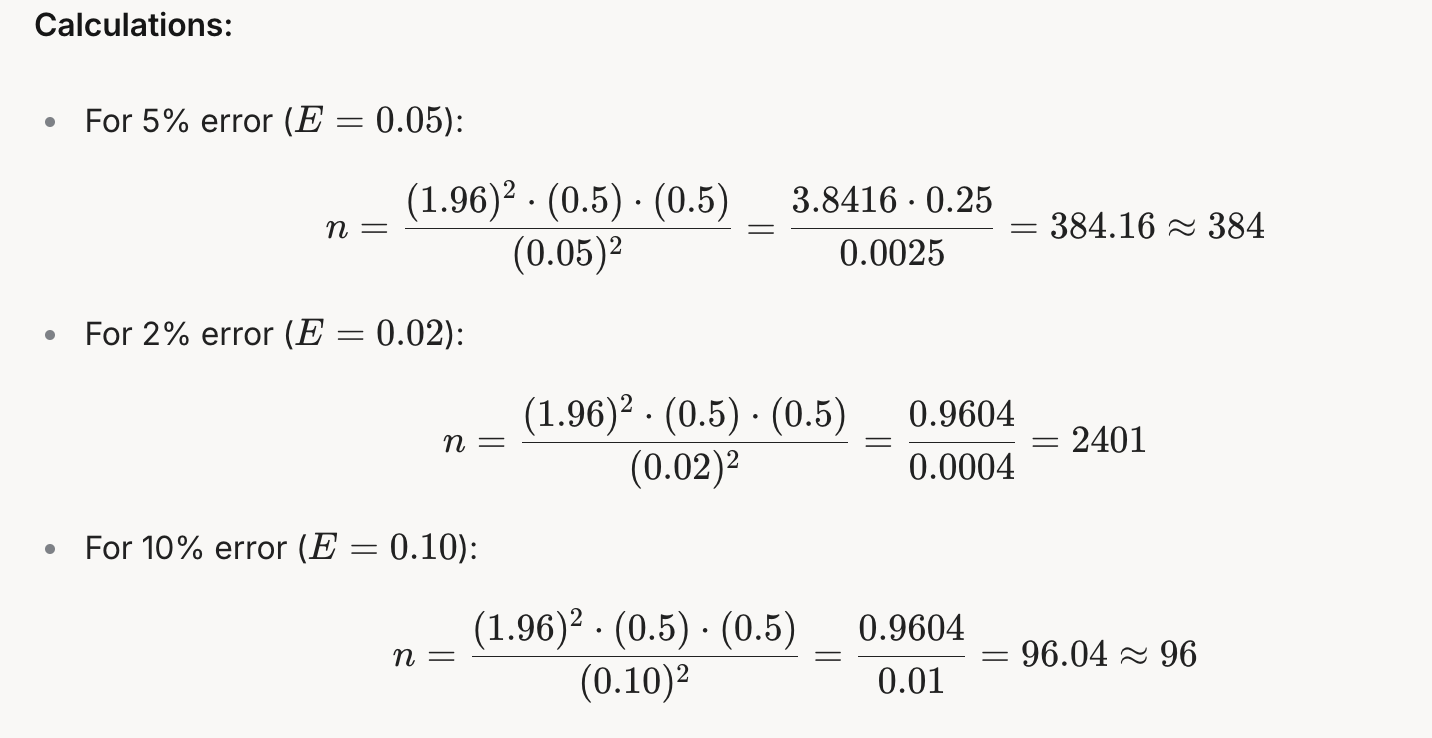
#### **3. Sample Size Calculation**

We use Cochran’s formula to determine the sample size for statistical significance, *assuming a large population (as N=243,228 is substantial):*



Where:

* Z = 1.96 (95% confidence level)
* p = 0.5 (maximum variability)
* E = Margin of error (5%, 2%, 10%)



**Sample Size Table:**

| Page Count | Total Documents | Sample (5% Error) | Sample (2% Error) | Sample (10% Error) |
| --- | --- | --- | --- | --- |
| 1 | 141,744 | 384 | 2,401 | 96 |
| 2 | 61,640 | 384 | 2,401 | 96 |
| 3 | 27,295 | 384 | 2,401 | 96 |
| 4 | 10,304 | 384 | 2,401 | 96 |
| 5 | 2,245 | 384 | 2,401 | 96 |

**Observation**: At a 5% error margin, ~384 documents per page count (total 1,920) ensure representativeness but may exceed annotation capacity. A 10% error margin (total 480) risks underrepresentation. These values are unadjusted for population size, as the populations are large enough to approximate an infinite population for practical purposes. However, adjustments could be made for smaller populations like Page 5 (2,245).

#### **4. Custom Sampling Distribution with Provided Sample Sizes**

We adopted a custom sampling approach, using predefined sample sizes per page count: [250, 200, 150, 100, 50] for Pages 1–5, totaling 750 documents:

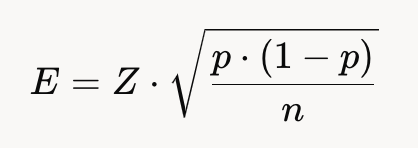
| Page Count | Total Documents | Proportion in Dataset | Custom Sample Size | Proportion in Sample |
| --- | --- | --- | --- | --- |
| 1 | 141,744 | 58.3% | 250 | 33.3% |
| 2 | 61,640 | 25.3% | 200 | 26.7% |
| 3 | 27,295 | 11.2% | 150 | 20.0% |
| 4 | 10,304 | 4.2% | 100 | 13.3% |
| 5 | 2,245 | 0.9% | 50 | 6.7% |
| **Total** | **243,228** | **100%** | **750** | **100%** |

**Split for Use (60:40):**

* Training (Fine-Tuning): 60% = 450 documents (150, 120, 90, 60, 30 for Pages 1–5).
* Testing (Evaluation): 40% = 300 documents (100, 80, 60, 40, 20 for Pages 1–5).

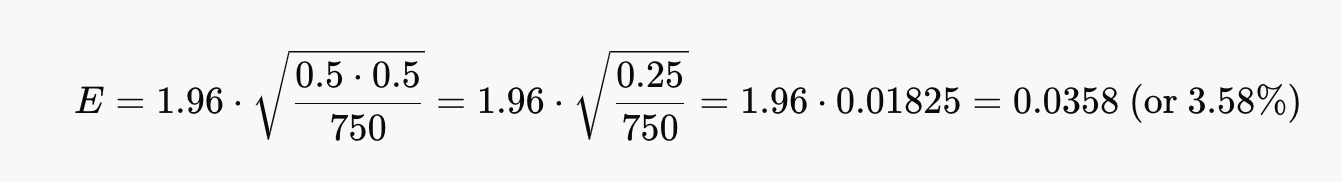
**Error Margin Analysis:**

* Overall Significance: For 750 documents from 243,228, the margin of error at 95% confidence is ±3.58%, calculated using the margin of error formula:



Where:

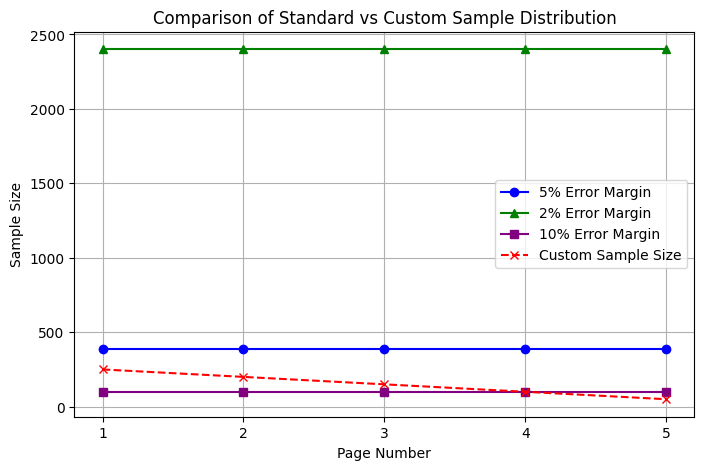
* E = Margin of error
* Z=1.96 (for 95% confidence level)
* p=0.5 (maximum variability)
* n=750 (total sample size)



This means the sample’s overall proportions are likely within ±3.58% of the true dataset values.

* Per-Page Error Margins: The table below shows the margin of error for each page count individually, based on its custom sample size, at 95% confidence, using the same formula:

| Page Count | Custom Sample Size | Proportion in Sample | 95% Confidence Margin of Error |
| --- | --- | --- | --- |
| 1 | 250 | 33.3% | ±6.2% |
| 2 | 200 | 26.7% | ±6.93% |
| 3 | 150 | 20.0% | ±8.0% |
| 4 | 100 | 13.3% | ±9.8% |
| 5 | 50 | 6.7% | ±13.86% |



* **Interpretation:** Pages with smaller sample sizes (e.g., Page 5 with 50 documents) have larger error margins (±13.86%), meaning we’re less precise about their proportions compared to pages with larger samples (e.g., Page 1 with 250 documents, ±6.2%). However, the overall sample’s error margin (±3.58%) balances these variations, ensuring reliable representation across all page lengths.
* **Comparison to Cochran’s Recommendations:**
  + Pages 1 and 2 (250 and 200) are below the 5% error threshold (~384) but exceed the 10% error threshold (96), indicating moderate representativeness.
  + Pages 3, 4, and 5 (150, 100, 50) are also below the 5% error threshold but improve representation compared to a proportional sample (84, 32, 7), prioritising longer documents.
* **Comparison to Proportional Sampling:** A proportional sample would yield 437, 190, 84, 32, 7 for Pages 1–5. The custom sample reduces Page 1 (250 vs. 437) and increases Pages 3–5 (150, 100, 50 vs. 84, 32, 7), intentionally overrepresenting longer documents to capture their complexity.

#### **5. Significance and Impact of the Custom Dataset**

* **Strategic Overrepresentation of Longer Documents:** The custom sample sizes [250, 200, 150, 100, 50] prioritise Pages 3–5 (20%, 13.3%, 6.7% in the sample vs. 11.2%, 4.2%, 0.9% in the dataset). This ensures the pipeline learns to handle multi-page documents, which often contain complex layouts and tables.
* **Balanced Representation with Focus:** While Page 1 is underrepresented (33.3% in sample vs. 58.3% in dataset), it remains the largest group (250 documents) with a tight error margin (±6.2%), providing sufficient training data for the most common case. Pages 3–5 have higher error margins (up to ±13.86% for Page 5), but their overrepresentation enhances the pipeline’s ability to generalise across varying document lengths, reducing the risk of failure on longer, less frequent documents.
* **Statistical Reliability:** The overall ±3.58% margin of error at 95% confidence confirms the 750-document sample is a reliable representation of the 243,228 documents with 1–5 pages. While per-page error margins vary (e.g., ±13.86% for Page 5), the overall sample’s precision ensures robust fine-tuning and evaluation, supporting accurate extraction of key information and tables.
* **Impact of 60:40 Split:** The 60:40 split (450 training, 300 testing) allocates 450 documents to training, providing a solid foundation for fine-tuning across all page lengths (e.g., 150 for Page 1, 30 for Page 5). The larger test set (300 documents) strengthens evaluation, offering greater confidence in the pipeline’s performance (e.g., 100 Page 1, 20 Page 5 in testing), even with higher error margins for smaller page counts. This split balances learning and validation effectively for the dataset size.
* **Improved Generalization:** Overrepresenting Pages 3–5 addresses potential challenges in extracting data from multi-page documents, where information like agreement numbers or tables may span pages. Despite the higher error margin for Page 5 (±13.86%), the 50 documents sampled provide valuable training and testing data for these rare cases, improving the pipeline’s performance on complex scenarios and enhancing its real-world applicability for motor finance operations.
* **Resource Efficiency:** The 750-document sample balances statistical significance with practicality, making annotation feasible while covering 98.2% of the dataset (1–5 pages). Excluding 6+ page documents aligns with their minimal relevance, saving resources without compromising core functionality.
* **Recommendation for Optimization:** The custom sample effectively prioritises longer documents, with per-page error margins providing insight into precision (e.g., Page 5’s ±13.86% indicates lower precision due to its small sample size). If precision for Page 5 is critical, consider increasing its sample size (e.g., to 75 by redistributing from Page 1). The larger test set (300) allows for thorough validation—consider using a subset for validation during fine-tuning to optimise performance further.

#### **6. Conclusion**

The custom sampling strategy, using [250, 200, 150, 100, 50] for Pages 1–5 with a 60:40 train/test split, ensures a 750-document dataset that is statistically significant (±3.58% error overall) and strategically balanced for motor finance use cases. By over representing longer documents, this approach enhances the pipeline’s ability to handle complex multi-page agreements, while per-page error margins (e.g., ±6.2% for Page 1, ±13.86% for Page 5) provide clarity on precision for each page count. The larger test set (300 documents) strengthens evaluation confidence. We recommend proceeding with this sample for fine-tuning and evaluation, with future testing on new real-world data to confirm performance. If longer documents (6+ pages) become relevant, a small supplemental sample can be added.