AI GENERATED VS HUMAN WRITTEN TEXT

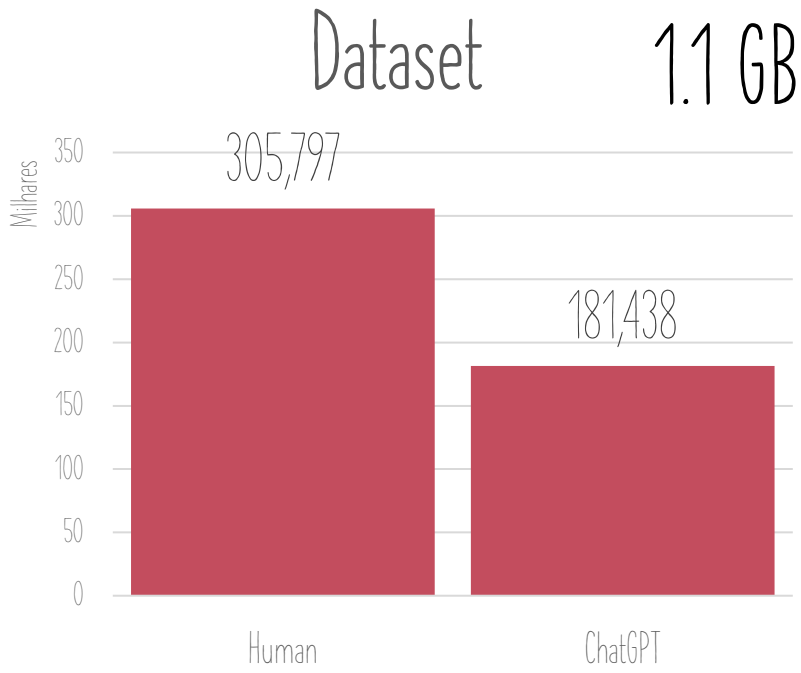
Teoria Algorítmica da Informação

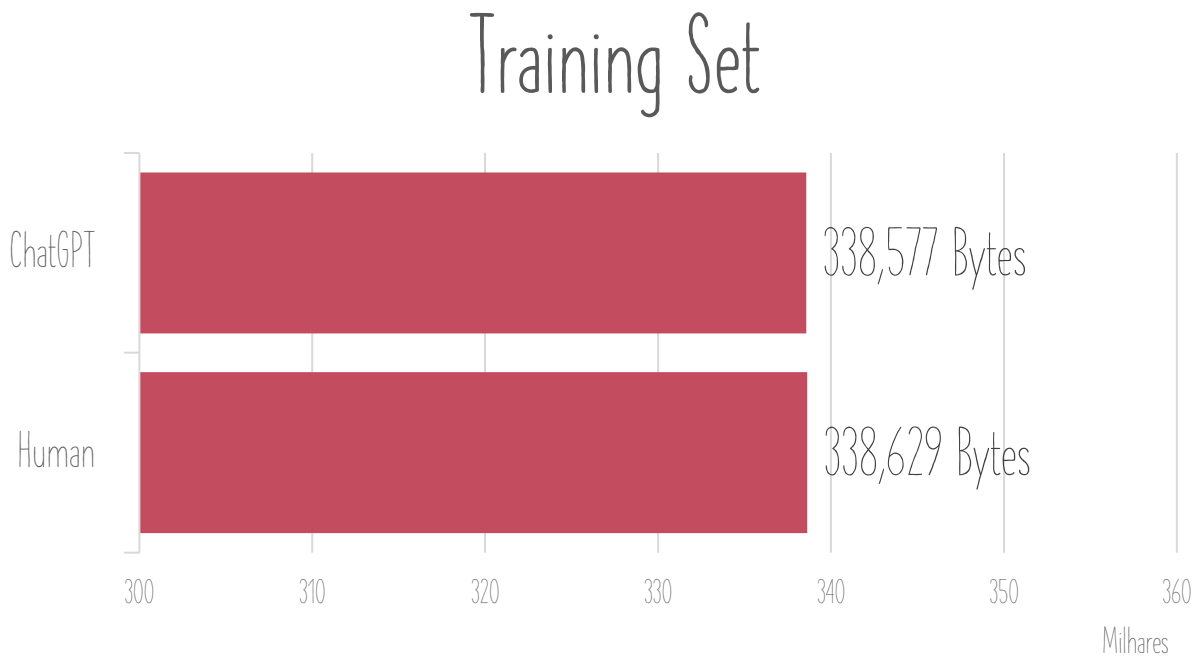
Gonçalo Silva - 103668 | Guilherme Antunes - 103600 | Pedro Rasinhas - 103541

THE PROPOSAL

- Classify texts as written by humans or generated by ChatGPT;
- Using only data compression as the sole feature to evaluate;
- Using finite context models to compress the texts for classification;
- Build an efficient solution in terms of speed and memory usage;

DATASET





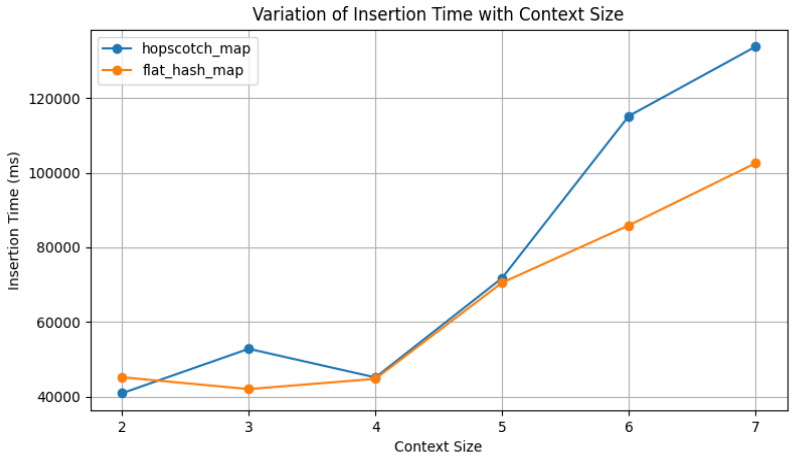
kaggle

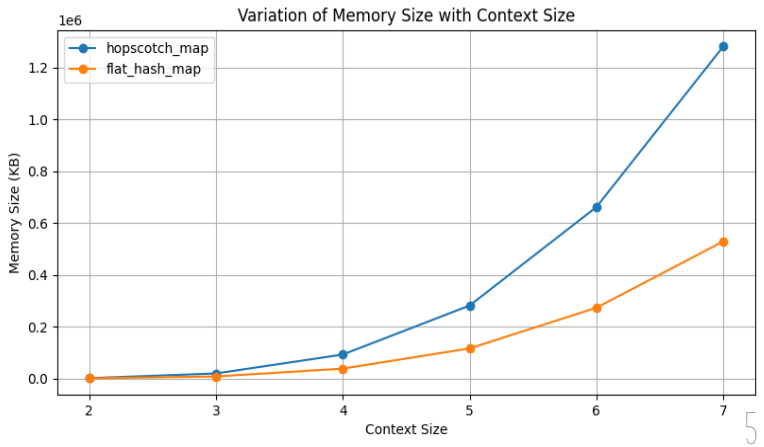
IMPLEMENTATION -> CHOICES AND DETAILS

- Implemented read buffers for the training sets;
- Replacement of strings by hashed strings in size_t;
- Usage of uints instead of ints for the counts;
- Usage of the '\0' char to separate files in the training set;
- Benchmarking different hash_maps to find the most efficient one;

IMPLEMENTATION -> TABLES

	unordered_map	flat_hash_map	sparse_hash_map	hopscotch_map
Tempo de Inserção (ms)	47002	20069	47142	18603
Tempo de Leitura (ms)	60488	35011	500246	29386
Memory Size (KB)	916	597	1343	1448



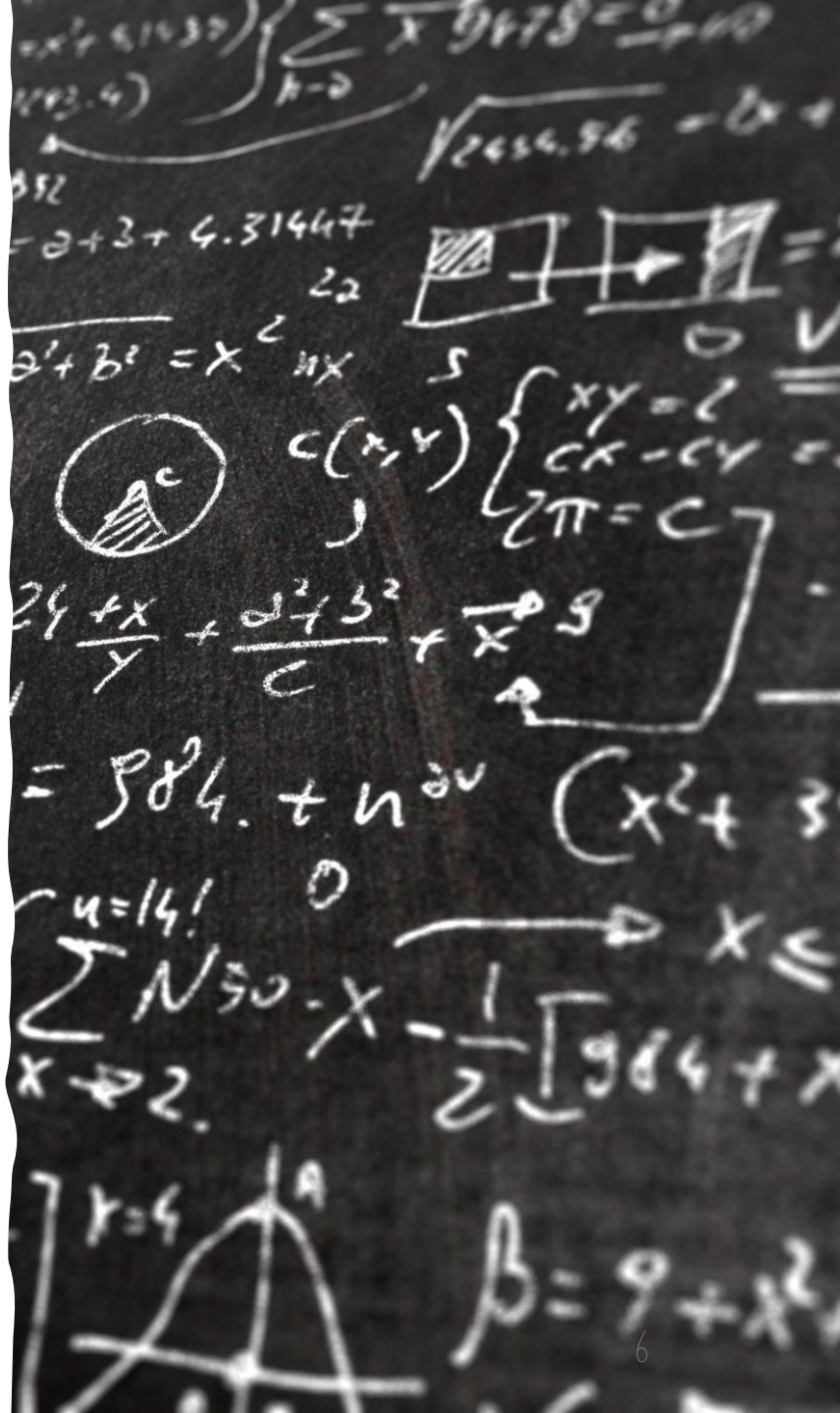


IMPLEMENTATION -> CALCULATIONS

$$P(y|x) = \frac{N(y|x) + \alpha}{\sum N(s|x) + \alpha \times |\Sigma|}$$

$$Memory = -\sum \log_2\big(P(y|x)\big)$$

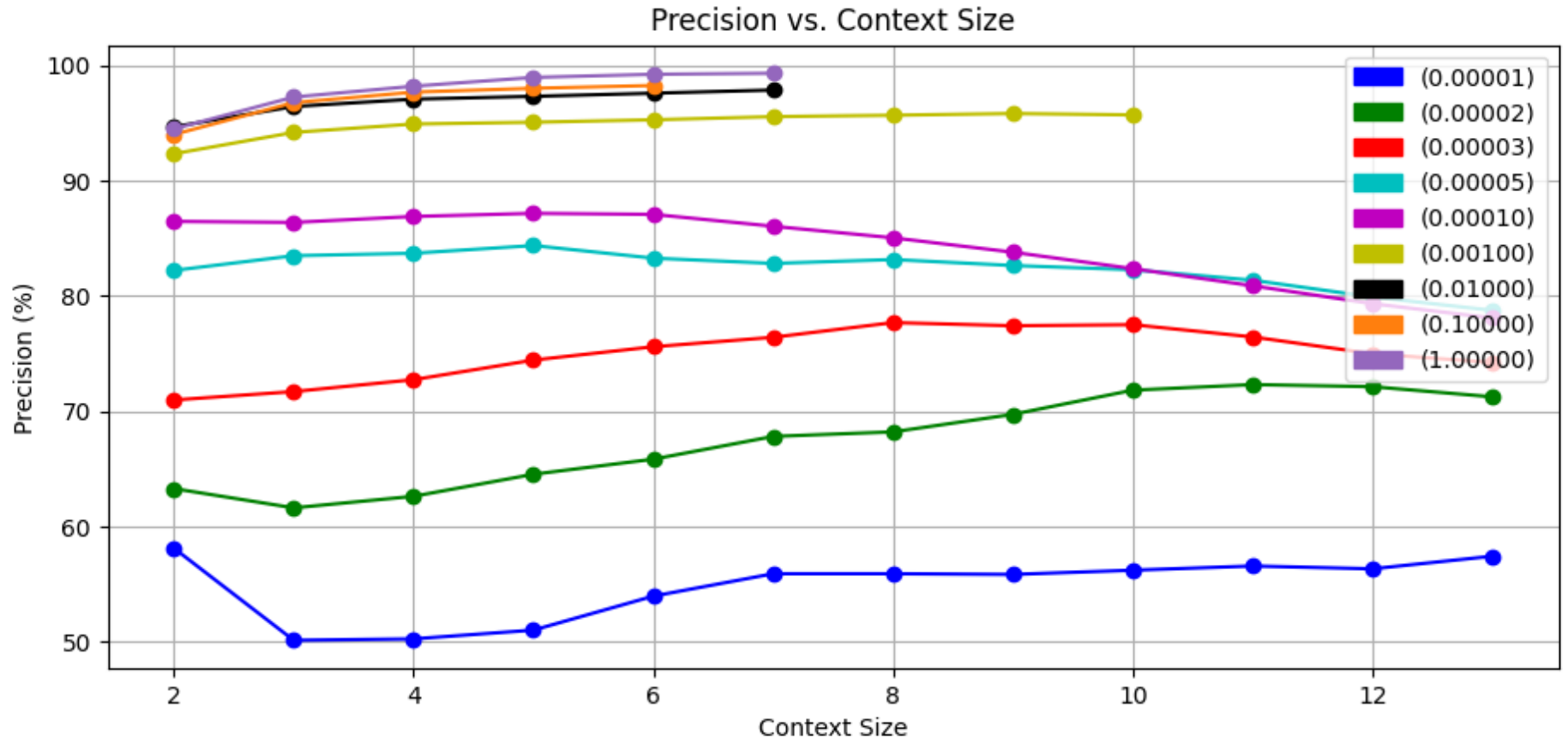
$$Precision = \frac{N_{hits}}{N_{hits} + N_{misses}}$$



EXTRA FEATURES

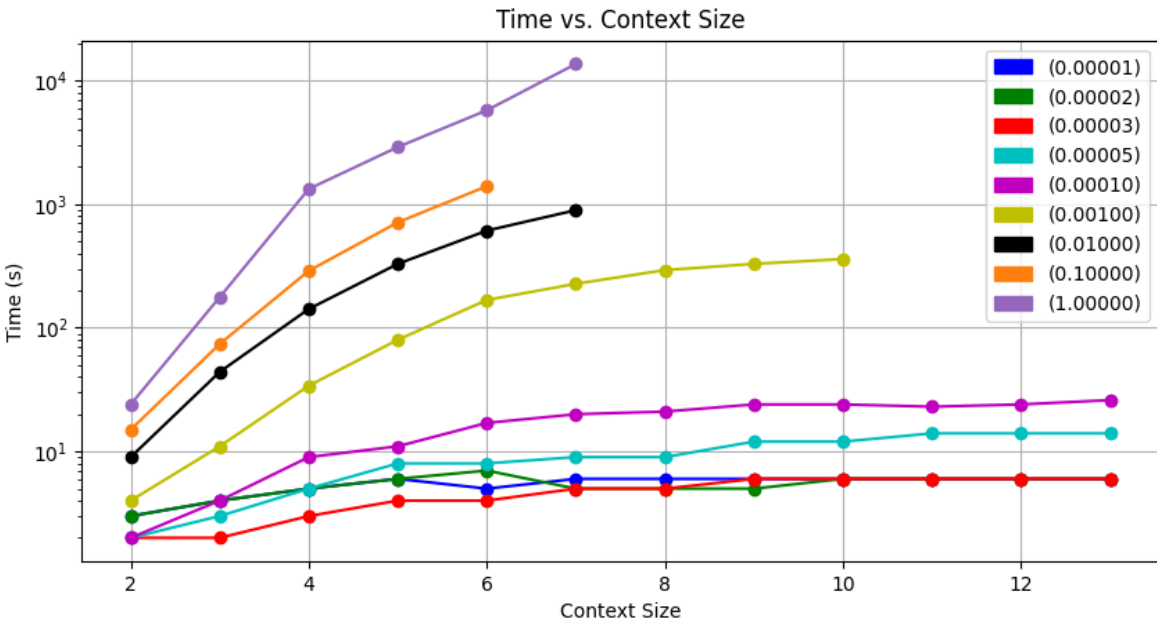
- Analyze entire directories;
- Analyze files iteratively (inserted 1 by 1);

RESULTS

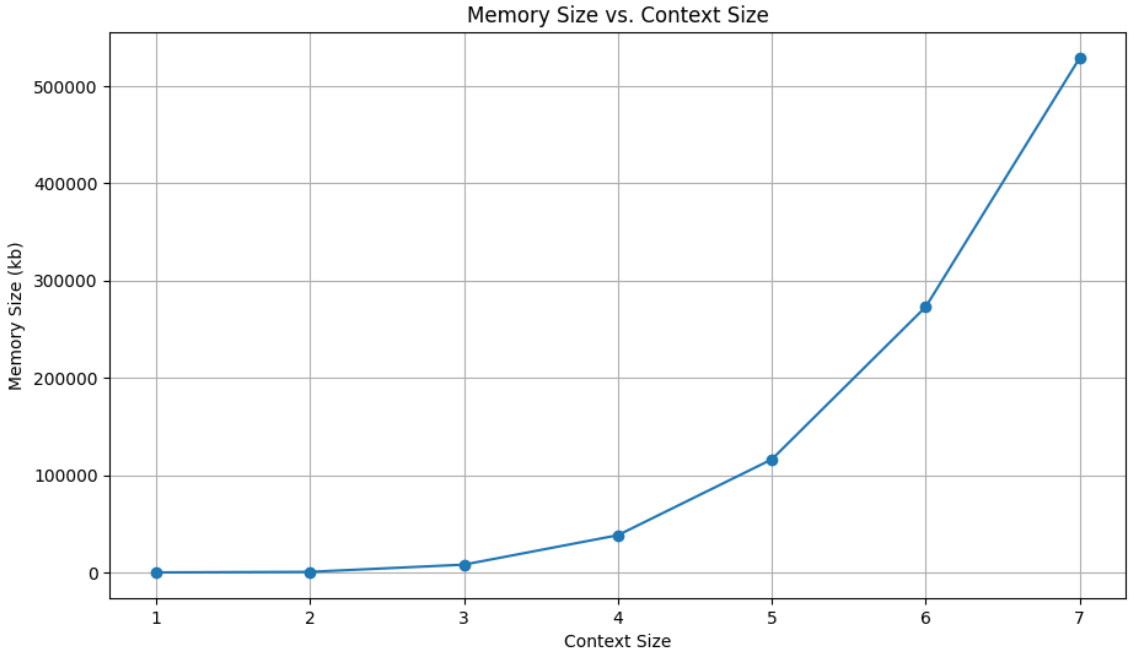


VARIATION OF MODEL ACCURACY WITH INCREASING CONTEXT SIZE FOR DIFFERENT TRAINING SETS

RESULTS

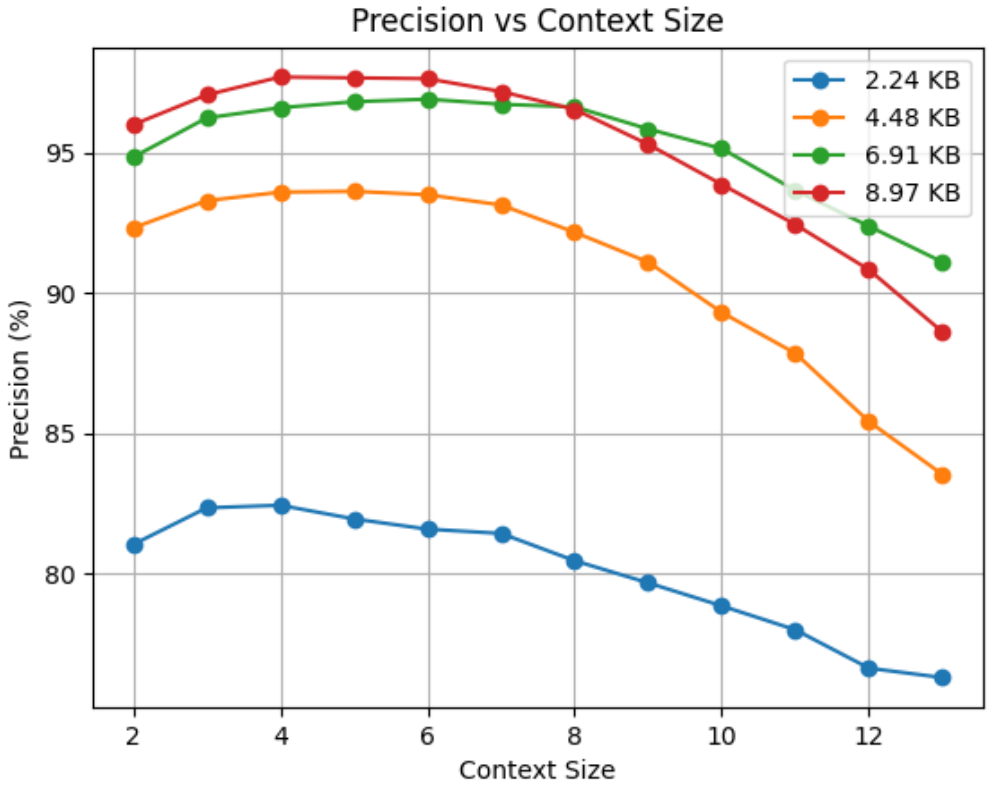


VARIATION OF READING TIME AS A FUNCTION OF INCREASING CONTEXT SIZE FOR DIFFERENT TRAINING SETS

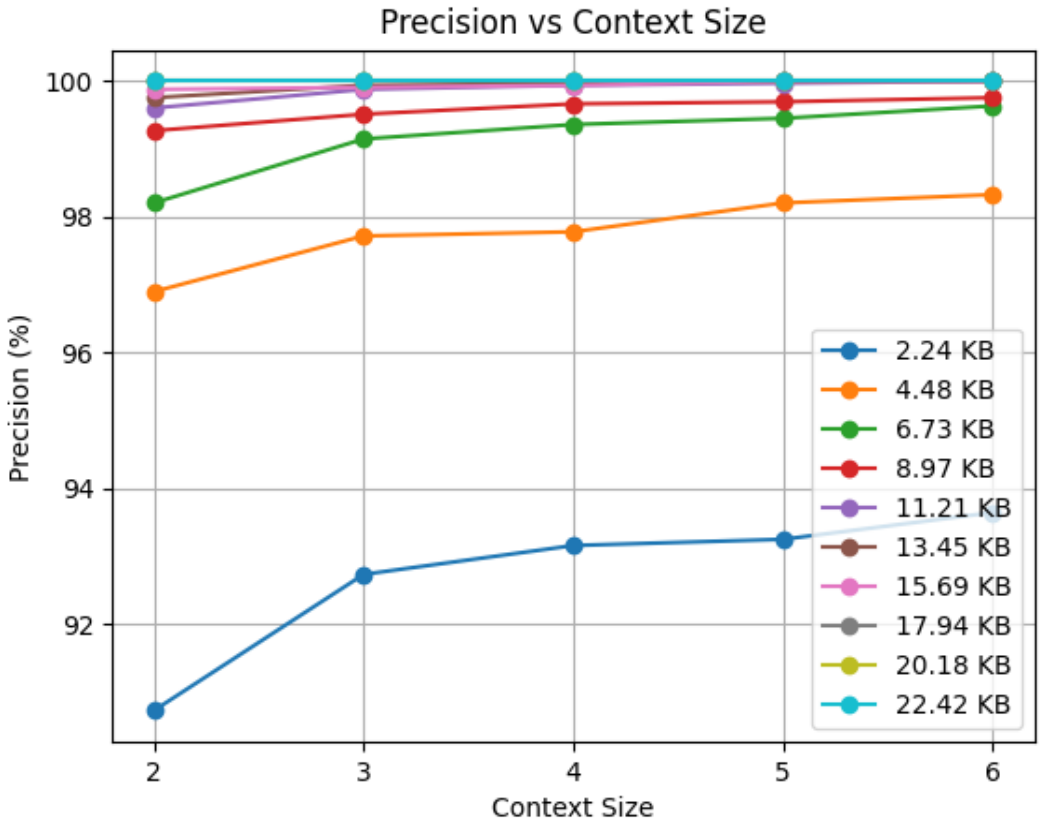


VARIATION OF OCCUPIED MEMORY AS A FUNCTION OF INCREASING CONTEXT SIZE FOR THE COMPLETE TRAINING SET

RESULTS



VARIATION OF ACCURACY AS A FUNCTION OF INCREASING CONTEXT SIZE FOR DIFFERENT TEST FILE SIZES AND SET 0.00001X THE TOTAL SIZE



VARIATION OF ACCURACY AS A FUNCTION OF INCREASING CONTEXT SIZE FOR DIFFERENT TEST FILE SIZES AND SET 0.001X THE TOTAL SIZE

CONCLUSIONS

- The larger the file, the higher the certainty in classification;
- The more training data is used, the larger context orders are supported;
- The larger the context, the easier it becomes to classify correctly;

THANK YOU!