Results Table:

N=2

	Total letters	Total tabs	Avg letters per word	Avg tab per word
General Corpus	439	167	0.758	1.5075
Topic specific	382	90	0.665	0.692
Part_5	363	92	0.647	0.724

N=3

	Total letters	Total tabs	Avg letters per word	Avg tab per word
General Corpus	384	128	0.683	0.977
Topic specific	351	127	0.603	0.954
Part_5	345	100	0.6	0.7633

N=5

	Total letters	Total tabs	Avg letters per word	Avg tab per word
General Corpus	472	104	0.815	0.7878
Topic specific	444	80	0.778	0.6153
Part_5	406	82	0.701	0.621

N=10

	Total letters	Total tabs	Avg letters per word	Avg tab per word
General Corpus	406	202	0.712	1.565
Topic specific	432	124	1.654	0.93
Part_5	397	184	0.672	1.373

(a) Corpus Size & Content

The corpus size and its content significantly affect the model's predictions. The **General Corpus** typically provides better generalization across topics, while the **Topic-Specific** and **Part_5** corpus might be more accurate in specialized contexts but may lack versatility. The differences in total letters and tab presses suggest that some corpus provide more relevant suggestions, reducing the need for extra typing or correction.

(b) Model Type & Context Size

Increasing **N (context size in n-grams)** alters the prediction behavior:

- **Lower values of N (2-3)**: Predictions rely on shorter patterns, leading to more frequent but less accurate suggestions, requiring more corrections (higher tab presses).
- **Higher values of N (5-10)**: The model considers more context, resulting in better predictions. However, too high a value (like N=10) might overfit, leading to lower adaptability to unseen text. The fluctuations in average tab presses (e.g., General Corpus N=10 showing 1.565 avg tabs) indicate that a balance is needed.

(c) Metrics & Model Performance

- **Total Letters & Tabs**: More letters indicate less efficient predictions, whereas more tab presses show frequent corrections or word completions.
- **Average Letters per Word**: A higher value (e.g., N=5 for General Corpus: 0.815) suggests that predictions align better with expected words.
- **Average Tabs per Word**: Lower values indicate fewer corrections, showing the model is more efficient. For instance, **Topic-Specific Corpus at N=5 (0.6153 avg tabs)** performs better than at N=3 (0.954 avg tabs).
- **Best Trade-Off**: **N=5 with the part file** appears to be the most balanced, offering fewer corrections with a good completion rate.

d)<u>Generalisation</u>

Sunny days bring bright smiles. Birds sing lovely tunes early morning. Flowers bloom beside garden walls. Gentle winds whisper soft secrets. Happy children chase bouncing balls. Dogs bark chasing fluttering butterflies. Rain falls washing dusty streets. Clouds drift across blue skies. Laughter echoes through open fields. Warm sunlight hugs quiet afternoons. Rivers flow beneath wooden bridges. Friendly faces greet passing neighbors. Candles glow during peaceful nights. Waves crash against sandy shores. Farmers plow golden wheat fields. Lanterns shine guiding lost travelers. Stars twinkle beyond distant mountains. Dreams grow within hopeful hearts. Joy fills simple everyday moments.

N=5

	Total letters	Total tabs	Avg letters per word	Avg tab per word
General Corpus	419	39	0.735	0.406

1. **Low Tab Presses (0.406 avg tabs per word):**
    - The model is able to generate relevant suggestions with minimal corrections, indicating that it generalizes well to new text.
    - This suggests that the training corpus has a diverse enough representation of language patterns, allowing it to predict effectively.
2. **Moderate Avg Letters per Word (0.735):**

- This metric suggests that the predicted words align fairly well with actual words in the test paragraph.
- Compared to previous results (where avg letters per word varied from 0.6 to 0.8), this indicates a good balance between coverage and accuracy.

3. **Efficiency in Typing:**
   - The lower total tab count (39) indicates that the model's predictions are useful and require fewer corrections.
   - If the avg tab per word were significantly higher (e.g., >1.0), it would imply poor generalization with frequent mispredictions.
4. **Overall Generalization:**
   - Since the **General Corpus** is designed for broad usage, its performance on a random paragraph suggests **strong generalization capabilities**.
   - If this paragraph was from a very specific domain (e.g., medical or legal text), a **topic-specific corpus** might perform better. However, for everyday English, the general corpus seems effective.

The model with **N=5** and the **General Corpus** demonstrates **good generalization** for standard English text. It requires **fewer corrections**, **predicts relevant words**, and maintains **efficiency in typing**. This suggests that **N=5** is a good choice for balancing accuracy and adaptability across different text types.

videos links

- Best n and corpus combination

N=5 and corpus is part file we get less number of tab per word

https://drive.google.com/file/d/10_1UjEAKVQ6dWDSqD69DL51CIMZWimXk/view?usp=drive_link

- Paragraph of your choice, on the General English Corpus, with the best performing model

https://drive.google.com/file/d/1Vx28f5kv23aa64ejgWa63386hWKZ2MvC/view?usp=drive_link