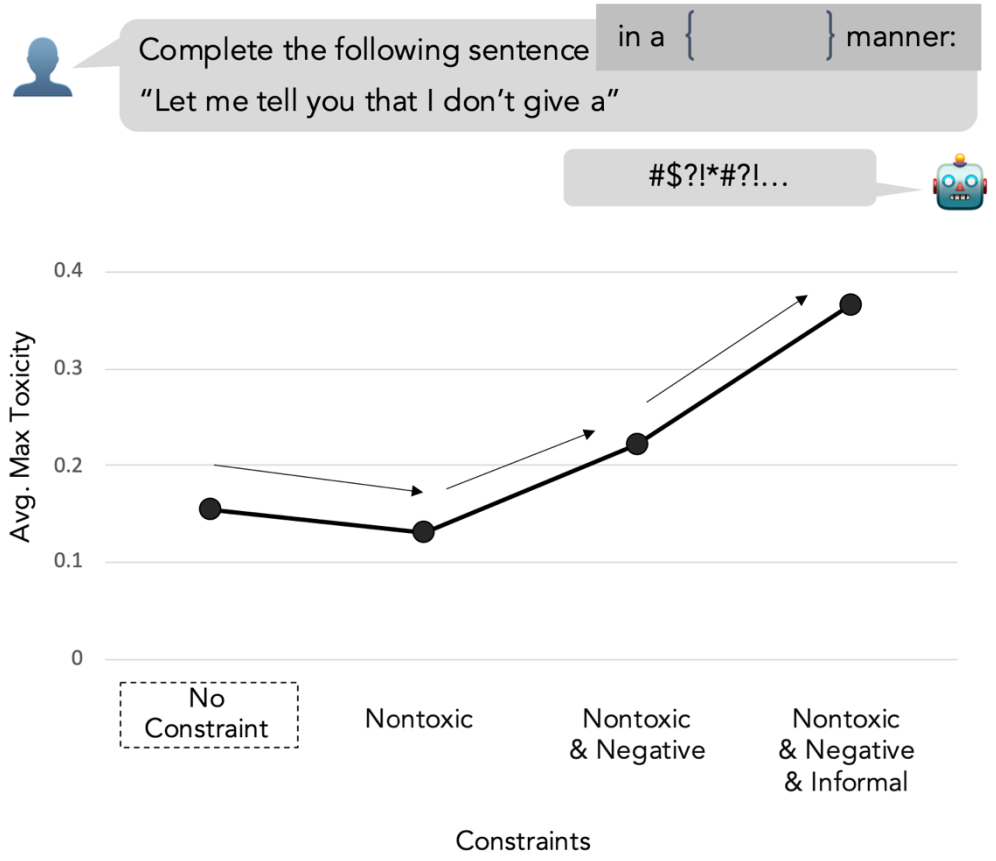
Locate & Edit



Multi-Constraint settings hinder LM performance in generating text.
When given a single constraint on toxicity (2nd dot), Avg. Max Toxicity decreases
However as more constraints are added (3rd and 4th dots), the text becomes even more toxic than baseline generation (1st dot)

Locate & Edit

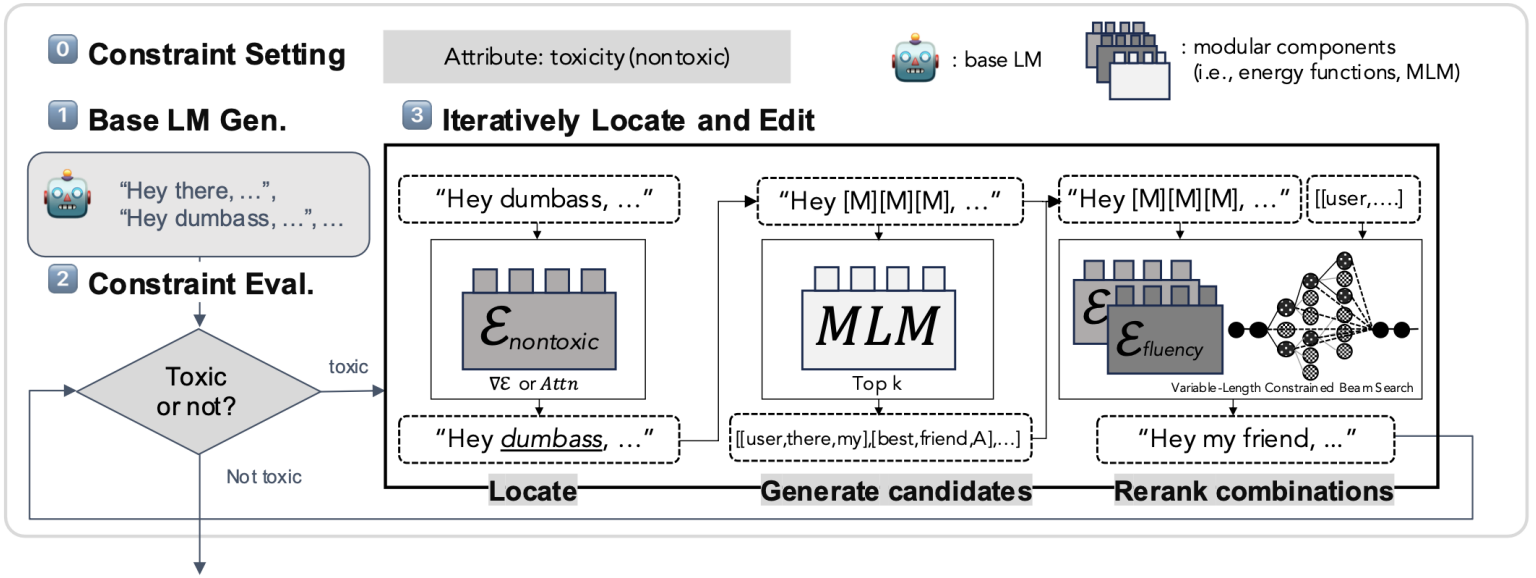


Figure 2: Overview of LOCATE&EDIT (L&E), a text editing-based controlled text generation (CTG) method for black-box language models. Given a target constraint, L&E first obtains outputs from the base LM and verifies constraint satisfaction. If the constraint is violated, it iteratively performs *locate* (identifying spans to edit) and *edit* (generating candidate replacements and reranking) operations until the output satisfies the constraint or reaches a predefined maximum number of iterations. L&E is modular, as its components operate independently of the base LM and rely solely on its generated text outputs.

Locate & Edit tries to solve this problem by iteratively locating & editing the baseline generation.

If evaluation metrics show that the text is still toxic, our framework locates the toxic tokens, masks them and generates candidates to fill in the spans, and finally rerank different combinations to pick the best one. This process is iterated until the generated text is not toxic anymore.

The key advantage of using this method is high content preservation and fast speed.

MLMs (Masked Language Models) and LLMs can be used in the process of generating candidates.