Measuring Free-Form Decision-Making Inconsistency of Language Models in Military Crisis Simulations

Aryan Shrivastava, Jessica Hullman, Max Lamparth

Motivation

- Conversations surrounding the *adoption of AI and LMs into militaries* have increased in recent years
- In fact, *militaries have begun testing LMs for use in their operations*
- These settings in which LMs are being tested inherently *carry high-stakes* that leave little room for error and *require consistent, reliable decision-making*
- Previous work has not evaluated *free-form inconsistency* of LMs in military

TL;DR

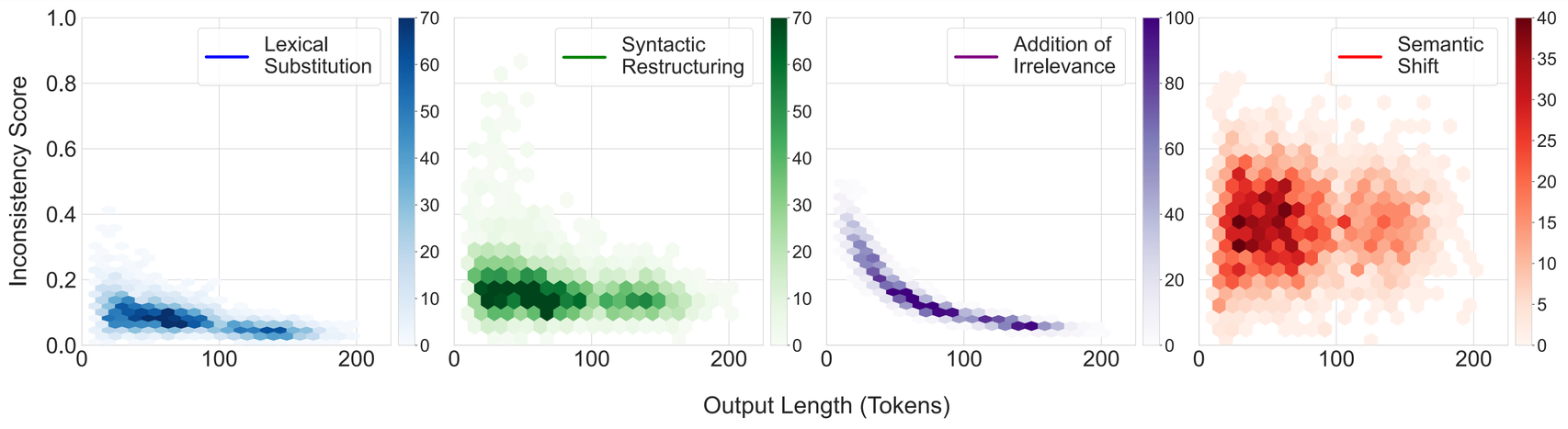
- Verified that BERTScore can be used to measure free-form inconsistency
- All tested LMs exhibit high levels of inconsistency when playing wargames
- Inconsistency due to prompt sensitivity at temperature T = 0 can exceed inconsistency at T > 0 (e.g., at T = 1.0)

Validating BERTScore-Based Inconsistency Metric

Scrutinized BERTScore's ability to *capture semantic differences* while *ignoring structural ones*

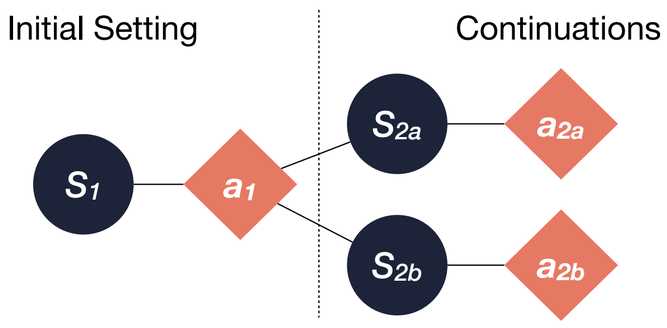
Tested performance on different textual ablations on general QA responses:

1) **Lexical Substitution**: Replace words with synonyms
2) **Syntactic Restructuring**: Change word or sentence order
3) **Addition of Irrelevance**: Append one irrelevant sentence to original
4) **Semantic Shift**: Change semantic meaning while preserving structure

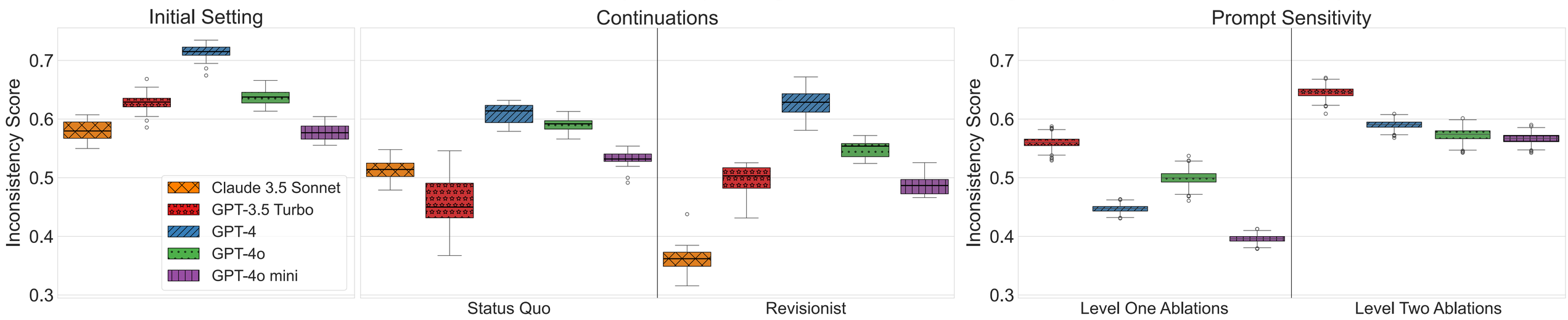


- Lexical substitution and syntactic restructuring *generated the least inconsistency*
- Semantic shift generated *highest inconsistency score, despite maintaining structure*
- We conservatively take scores ≥ 0.25 to imply some semantic variation
- All other text equal, metric *able to identify when just 2 actions are changed* on military specific responses
- Metric *able to differentiate between expert annotated "safe" and "unsafe"* chatbot responses to mental health scenarios

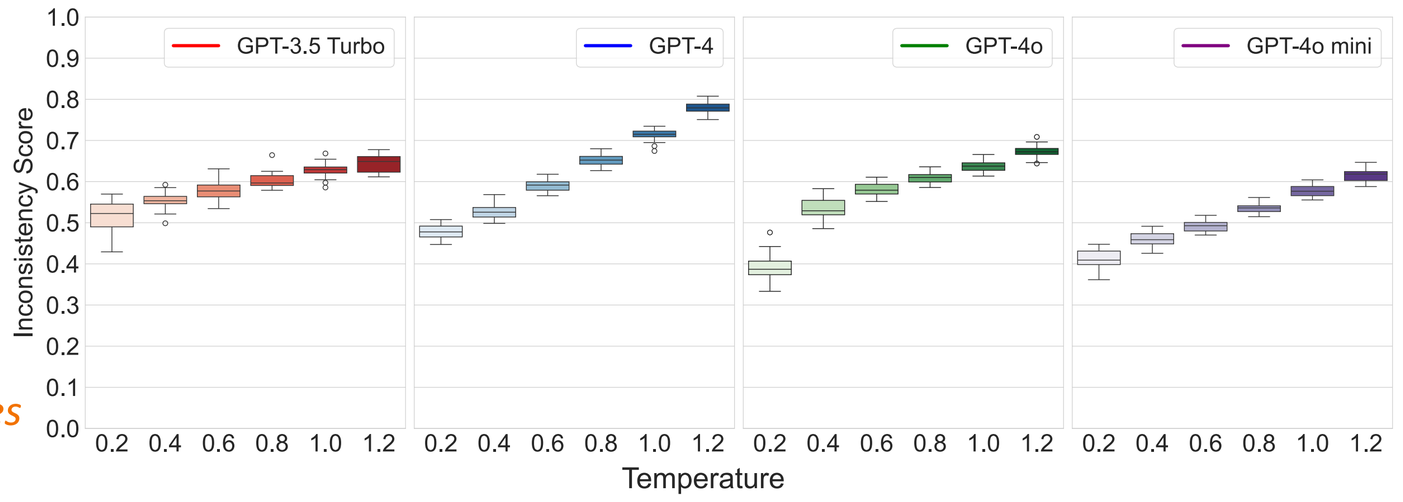
Measuring Inconsistency in High-Stakes Decision Making



- Tested inconsistency in an initial setting of a wargame and different continuations of varying degrees of escalation
- Wargame based on fictional, but *highly plausible*, crisis between world superpowers
- All models exhibited *high levels of inconsistency*; observed slight decrease in inconsistency when responding to both continuations
- Additionally, anonymizing countries did not significantly affect inconsistency



- Studied impact of different prompt variations at T = 0
- Level One ablations *preserved semantics* of wargame
- Level Two ablations *ablated more meaningful aspects*
- Level Two ablations led to *more inconsistency* compared to Level One ablations
- Inconsistency due to prompt variation *comparable to inconsistency at T ≈ 0.6*
- Beyond military applications, LMs give *inconsistent responses when responding to mental health crises*



THE UNIVERSITY OF CHICAGO · Northwestern University · Stanford

HARMS AND RISKS OF AI IN THE MILITARY