
expoQA®26
MADRID 26th, 27th & 28th May
expoqa.eu

Hello! 👋

I’m Jalpa Soni

Senior Data Scientist at Capitole

Welcome to our Talk:

The Rabbit Hole of Grammar: **LLMs on Trial with QA Adventures**

The Problem?

Traditional QA

- ✓ *Brittle scripts* break under dynamic UI changes
- ✓ *Hard-coded assertions* fail
- ✓ No reliable test oracles for *semantic correctness*
- ✓ *Multilingual testing multiplies complexity*

```
# Pseudo-Selenium test

driver.find_element(By.ID, "submit-btn").click()

# Hard-coded assertion

message = driver.find_element(By.ID, "confirmation").text
assert message == "Your request has been submitted successfully."
```

Traditional QA

- ✓ *Brittle scripts* break under dynamic UI changes
- ✓ *Hard-coded assertions* fail
- ✓ No reliable test oracles for *semantic correctness*
- ✓ *Multilingual testing multiplies complexity*

```
# Pseudo-Selenium test

driver.find_element(By.ID, "submit-btn").click()

# Hard-coded assertion

message = driver.find_element(By.ID, "confirmation").text
assert message == "Your request has been submitted successfully."
```

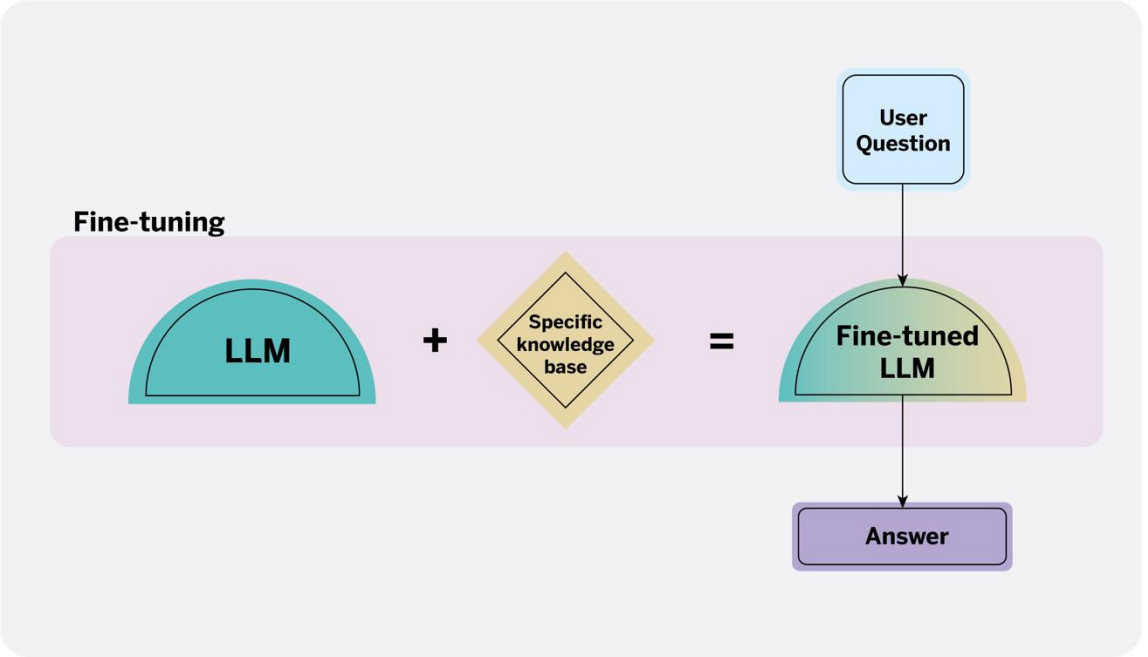
Enter LLMs,

- ✓ Outputs are probabilistic, not deterministic
- ✓ *Meaning matters more than exact string match*
- ✓ Need for semantic, not syntactic validation
- ✓ Opportunity: *LLMs can evaluate other LLMs*

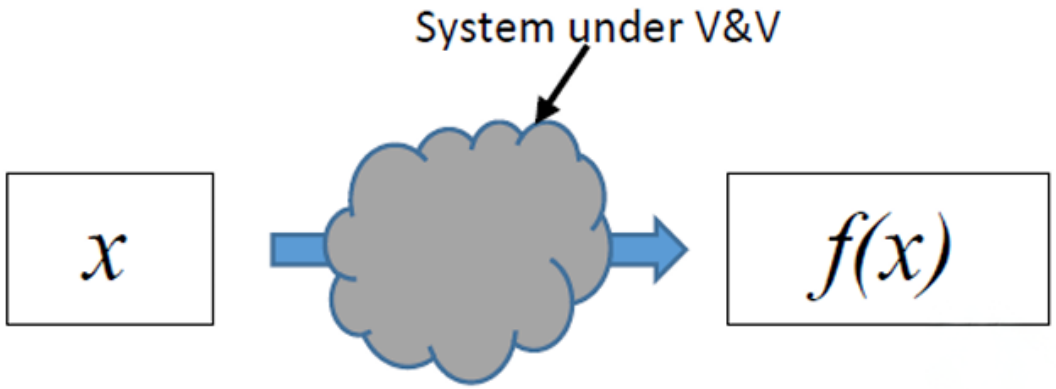
The Solution

Automated LLM Testing Pipeline

- ✓ End-to-end functional testing for web apps
- ✓ Generate, validate, and mutate test cases
- ✓ Metamorphic testing ensures consistency
  - If we **permute** the values in the text file, the results should stay the same
  - If we **multiply** each score by 10, the final results should all be multiplied by 10 as well
- ✓ Universal Grammar (UG) provides theoretical backbone

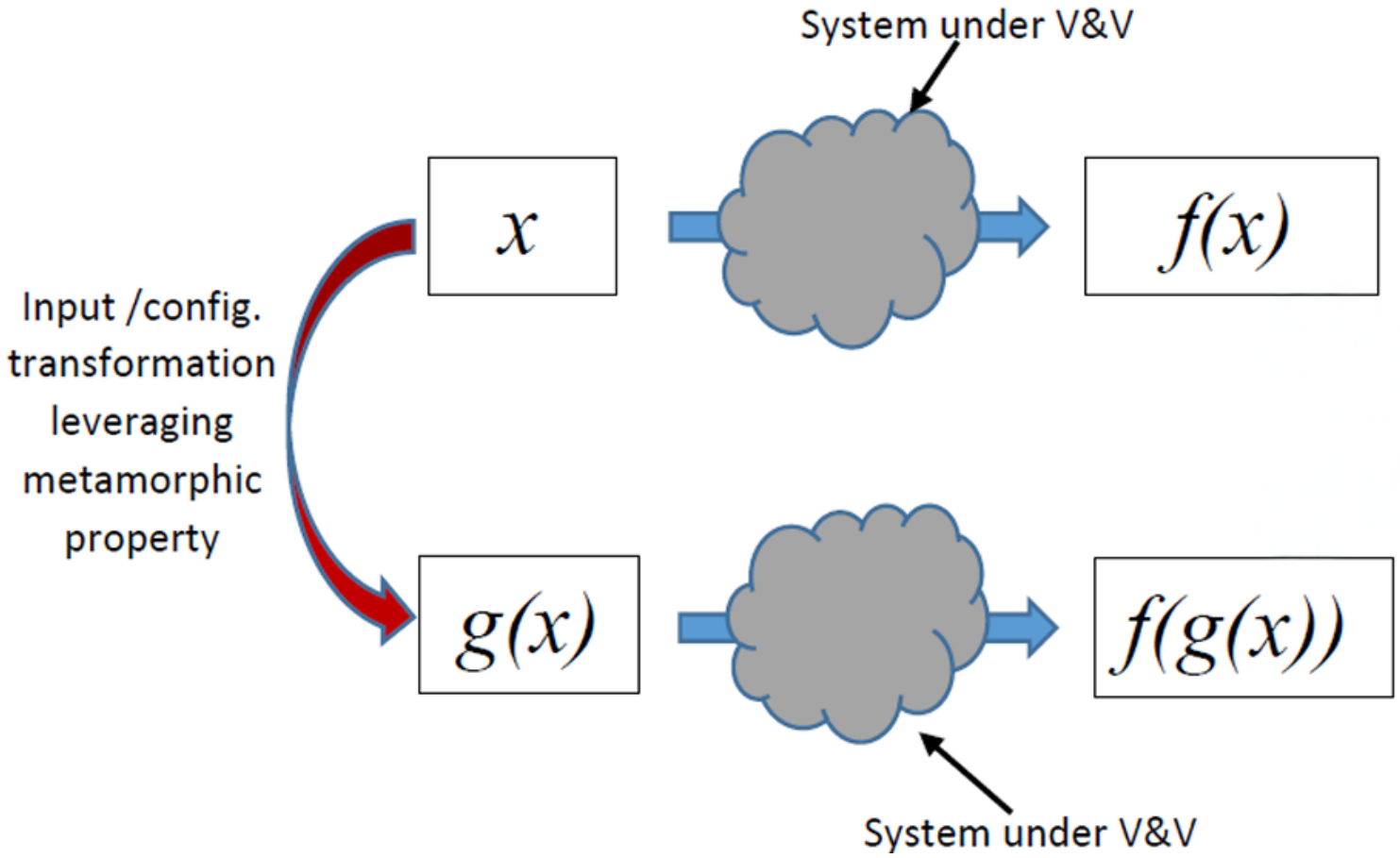


Metamorphic Testing

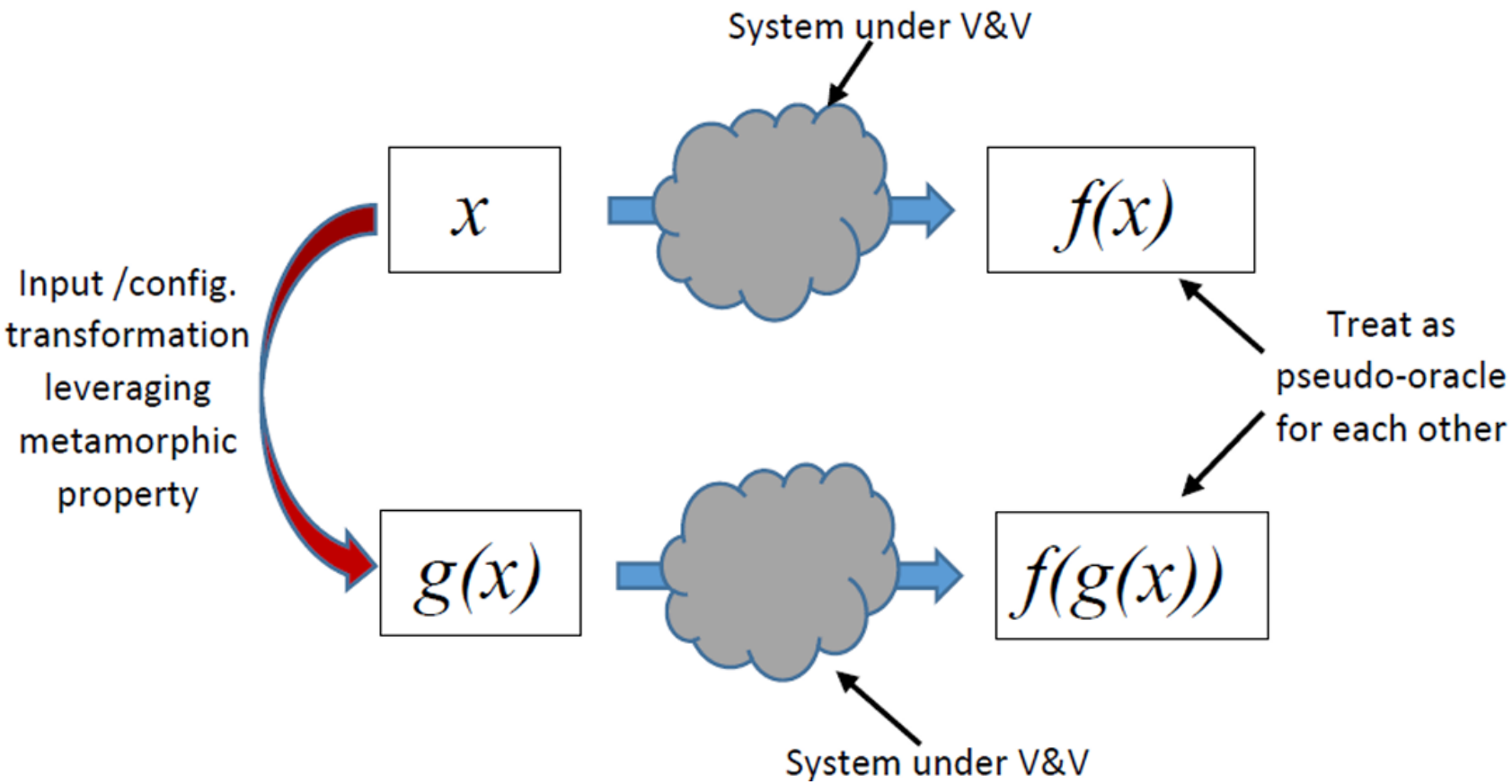


Raunak & Olsen (2015). Simulation Validation Using Metamorphic Testing

Metamorphic Testing



Metamorphic Testing



Theoretical Backbone

Connecting linguistic theory to software testing

Universal Grammar – In a Nutshell

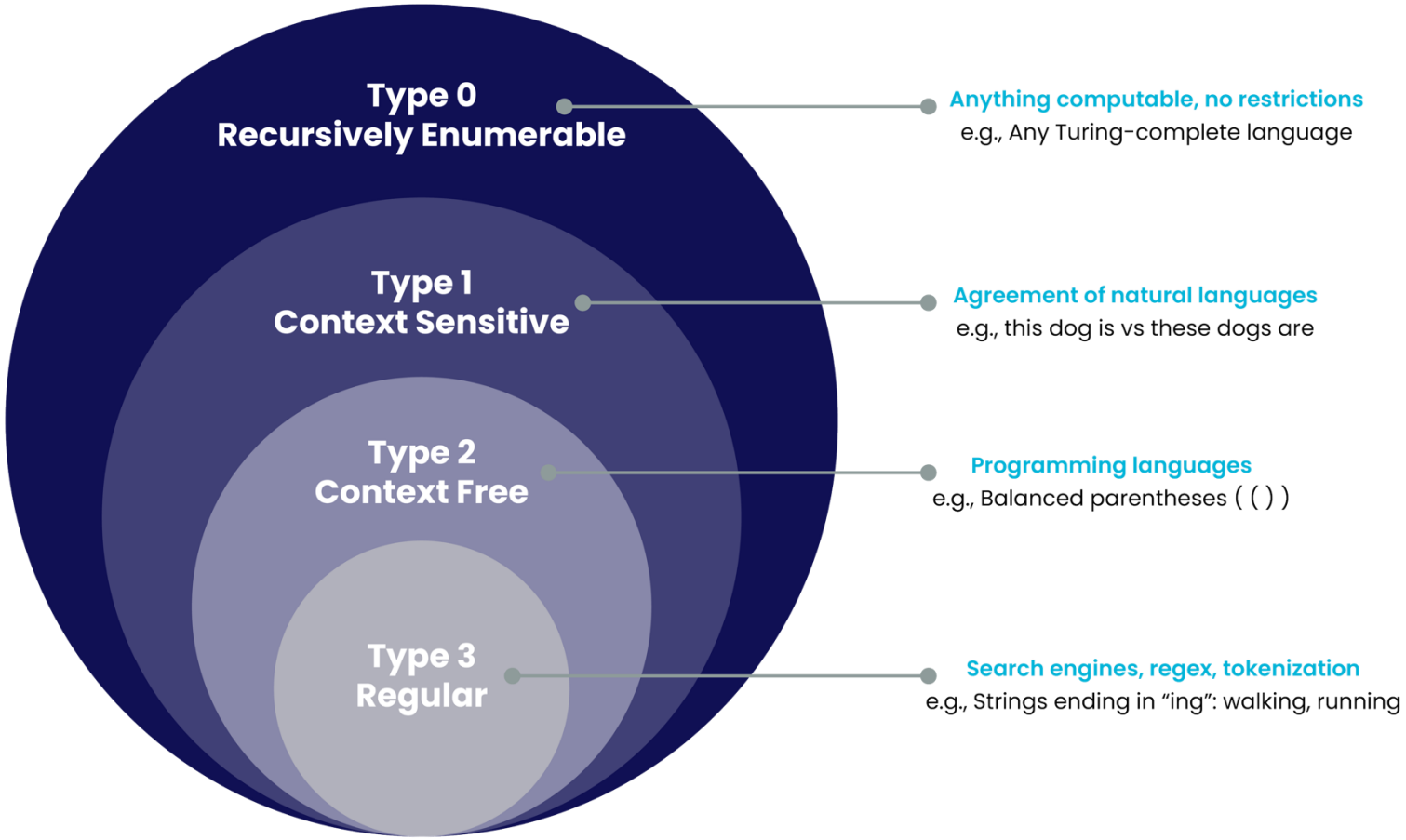
Noam Chomsky:

- ✓ The human brain contains a limited ***set of rules*** for organizing language
- ✓ All languages have a ***common structural basis***
- ✓ This set of rules is known as ***Universal Grammar***

Universal Grammar – In a Nutshell

Noam Chomsky:

- ✓ The human brain contains a limited ***set of rules*** for organizing language
- ✓ All languages have a ***common structural basis***
- ✓ This set of rules is known as ***Universal Grammar***

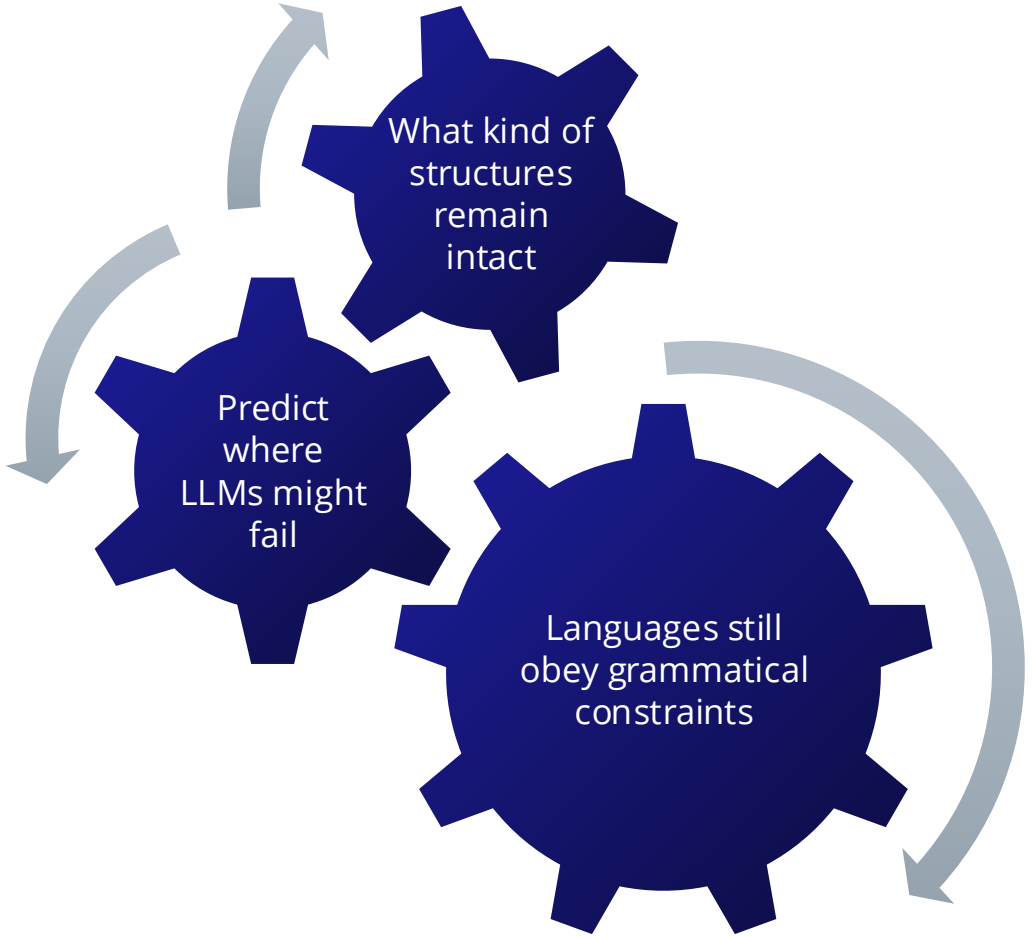


Universal Grammar – Why do we care?

LLMs don’t explicitly use grammars…

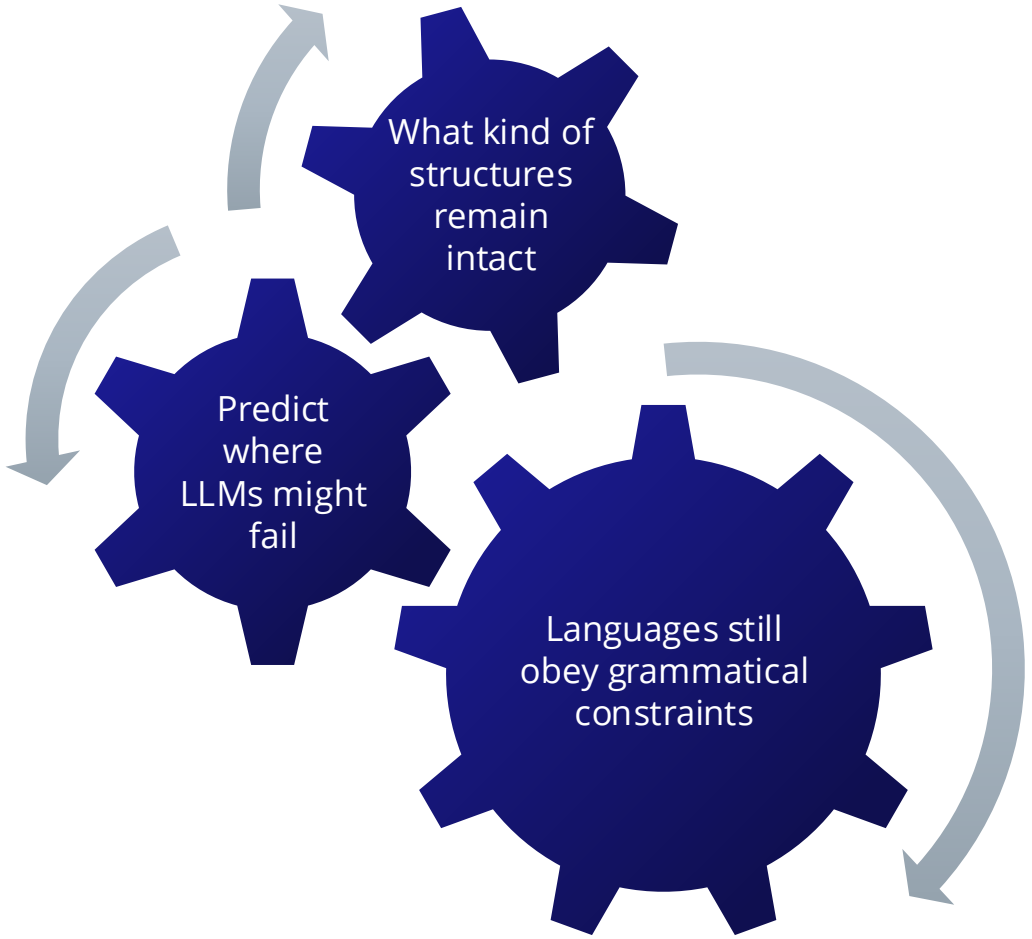
Universal Grammar – Why do we care?

LLMs don’t explicitly use grammars…



Universal Grammar – Why do we care?

LLMs don’t explicitly use grammars…



Using UG can help to highlight,

Language agnostic validation

semantic fidelity across translations.

Universal Grammar + Metamorphic Relations

MRs	Base	Transform	Expected	Check
1	She is happy	She is joyful	Ella está alegre → Ella está contenta	Reflect the synonym shift Not identical wording but equivalent meaning
2	He reads books	He reads books *in the evening*	Él lee libros por la tarde	Should include the temporal modifier
3	The tall man with a red hat walked slowly	The man ~~*with a red hat*~~ walked slowly	El hombre caminó despacio	Should reflect reduced structure without residual modifiers
4	Yesterday, she cooked dinner	She cooked dinner *yesterday*	Ayer, ella cocinó la cena	Should preserve meaning regardless of English word order
5	The boy runs	The boy*s* run~~*s*~~	El niño corre → Los niños corren	Should reflect correct gender and plural agreement

Universal Grammar + Mutations

Type	Mutation	Explanation
Lexical	Random Typos	Spelling errors at the word level
Lexical	Realistic Typos	Spelling errors mimicking human typing mistakes
Lexical	Shuffle Characters	Rearranging letters within a word
Lexical	Add Characters	Inserting extra letters into a word
Lexical	Remove Characters	Deleting selected letters from a word (e.g. remove all ‘e’ from a word)
Lexical	Delete Characters	Deleting random characters from a word

Universal Grammar + Mutations

Type	Mutation	Explanation
Lexical	Random Typos	Spelling errors at the word level
Lexical	Realistic Typos	Spelling errors mimicking human typing mistakes
Lexical	Shuffle Characters	Rearranging letters within a word
Lexical	Add Characters	Inserting extra letters into a word
Lexical	Remove Characters	Deleting selected letters from a word (e.g. remove all ‘e’ from a word)
Lexical	Delete Characters	Deleting random characters from a word
Semantic	Synonym Replacement	Substituting a word with another that changes meaning nuance
Semantic	Gender Swap	Alters meaning by changing gendered terms (e.g., “he” → “she”)
Semantic	Tense Shift	Changes temporal meaning (past *vs* present *vs* future)
Semantic	Quantifier Flip	Alters logical meaning (e.g., “some” → “all”)

Universal Grammar + Mutations

Type	Mutation	Explanation
Lexical	Random Typos	Spelling errors at the word level
Lexical	Realistic Typos	Spelling errors mimicking human typing mistakes
Lexical	Shuffle Characters	Rearranging letters within a word
Lexical	Add Characters	Inserting extra letters into a word
Lexical	Remove Characters	Deleting selected letters from a word (e.g. remove all 'e' from a word)
Lexical	Delete Characters	Deleting random characters from a word
Semantic	Synonym Replacement	Substituting a word with another that changes meaning nuance
Semantic	Gender Swap	Alters meaning by changing gendered terms (e.g., "he" → "she")
Semantic	Tense Shift	Changes temporal meaning (past *vs* present *vs* future)
Semantic	Quantifier Flip	Alters logical meaning (e.g., "some" → "all")
Syntactic	Word Shuffle	Rearranging word order affects sentence structure
Syntactic	Passive to Active	Voice transformation changes sentence structure
Syntactic	Cleft sentence	Restructuring into cleft form ("It is X that …")
Syntactic	Relative Clause	Adding or modifying relative clauses changes structure

Universal Grammar + Mutations

Type	Mutation	Explanation
Lexical	Random Typos	Spelling errors at the word level
Lexical	Realistic Typos	Spelling errors mimicking human typing mistakes
Lexical	Shuffle Characters	Rearranging letters within a word
Lexical	Add Characters	Inserting extra letters into a word
Lexical	Remove Characters	Deleting selected letters from a word (e.g. remove all 'e' from a word)
Lexical	Delete Characters	Deleting random characters from a word
Semantic	Synonym Replacement	Substituting a word with another that changes meaning nuance
Semantic	Gender Swap	Alters meaning by changing gendered terms (e.g., "he" → "she")
Semantic	Tense Shift	Changes temporal meaning (past *vs* present *vs* future)
Semantic	Quantifier Flip	Alters logical meaning (e.g., "some" → "all")
Syntactic	Word Shuffle	Rearranging word order affects sentence structure
Syntactic	Passive to Active	Voice transformation changes sentence structure
Syntactic	Cleft sentence	Restructuring into cleft form ("It is X that …")
Syntactic	Relative Clause	Adding or modifying relative clauses changes structure

Surface form

Meaning change

Structure change

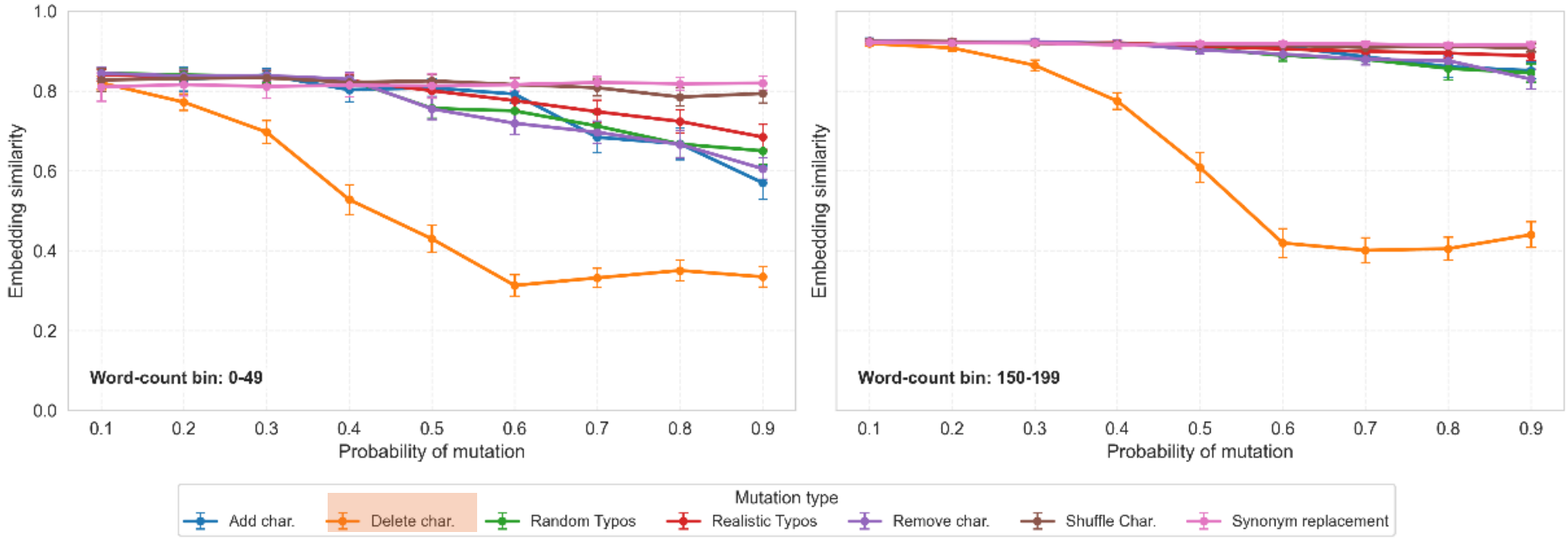
Results

Mutations Metrics

Metric	What it measures	In Chomskean mutation testing
Embedding Similarity $Sim_{cos}(x, y) = \frac{e_x \times e_y}{\|\|e_x\|\| \, \|\|e_y\|\|}$	Semantic closeness between original and mutated text	• Detects meaning drift when syntactic mutations are applied • Shows whether deep structure is preserved

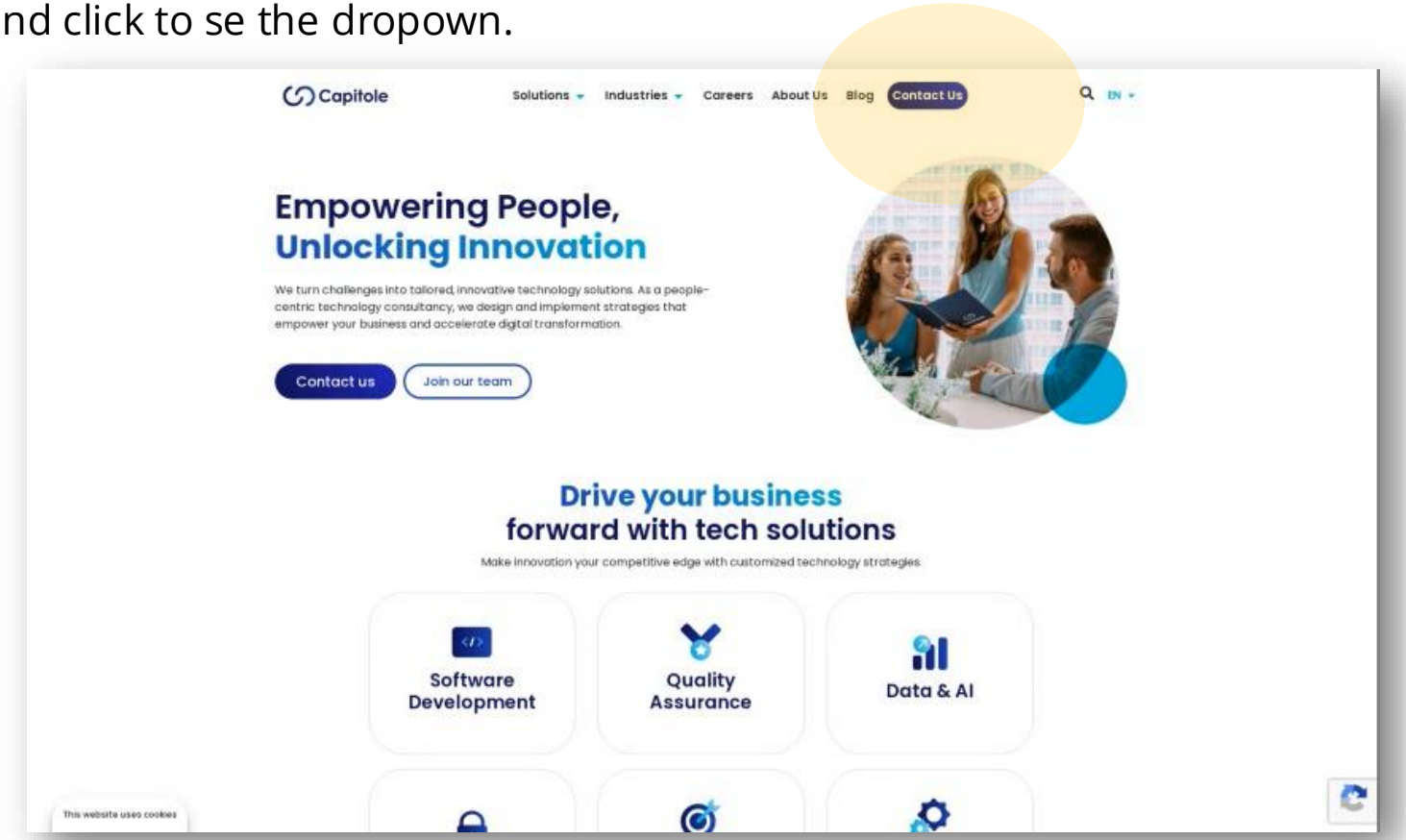
Mutations Metrics

Metric	What it measures	In Chomskean mutation testing
Embedding Similarity $Sim_{cos}(x,y) = \frac{e_x \times e_y}{\|\|e_x\|\| \, \|\|e_y\|\|}$	Semantic closeness between original and mutated text	• Detects meaning drift when syntactic mutations are applied • Shows whether deep structure is preserved



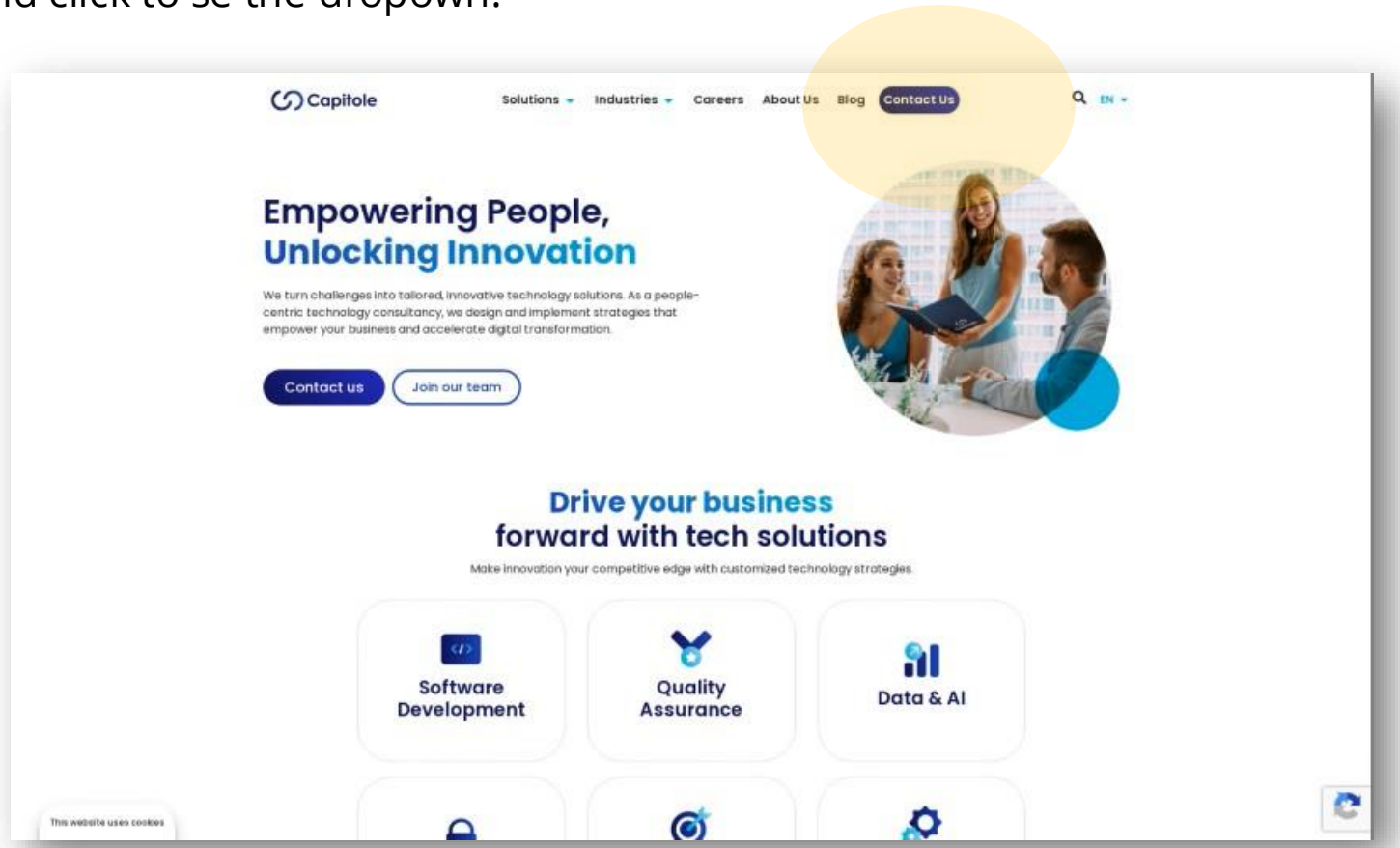
Example 1: No Mutation

Navigte to https://www.capitole-consulting.com/. Highlght the "Contct us" buton and click to se the dropown.



Example 2: 10% delete character

Navigte to https://www.capitole-consulting.com/. Highlght the "Contct us" buton and click to se the dropown.



Example 2: 70% delete character

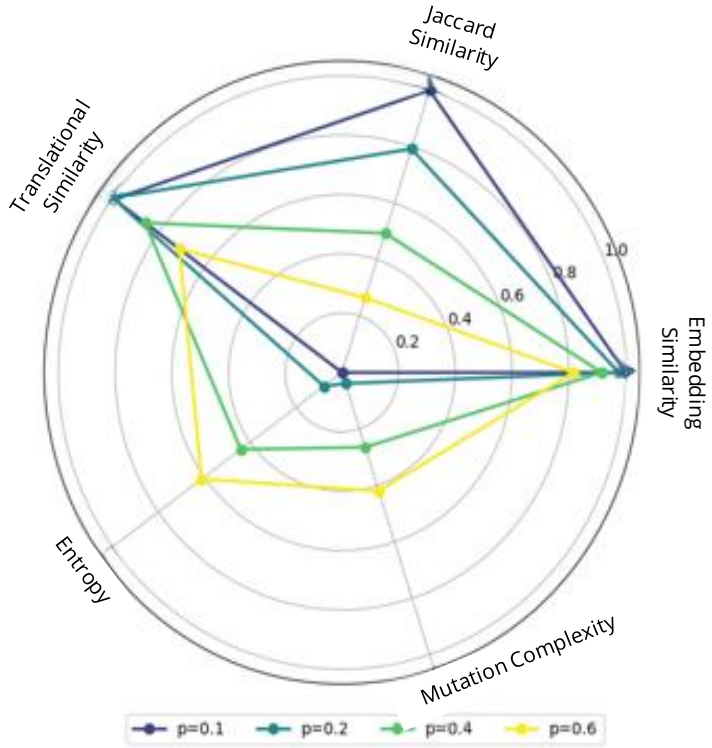
Go https://www.capitole-consulting.com/. hig t "cont us" bt cl t se drp.

Mutations Metrics

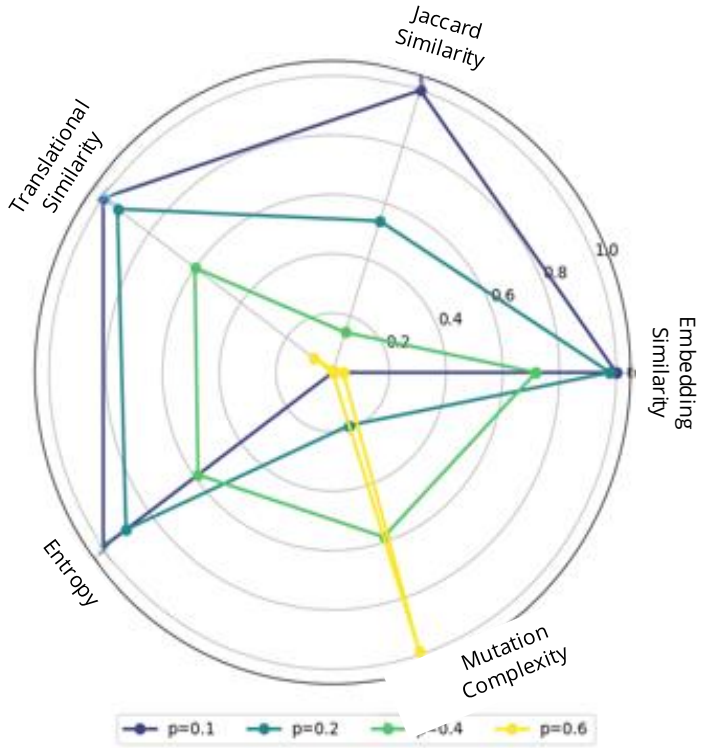
Metric	What it measures	In Chomskean mutation testing
Embedding Similarity $Sim_{cos}(x,y) = \frac{e_x \times e_y}{\|\|e_x\|\| \, \|\|e_y\|\|}$	Semantic closeness between original and mutated text	• Detects meaning drift when syntactic mutations are applied • Shows whether deep structure is preserved
Jaccard Similarity $J(A,B) = \frac{\|A \cap B\|}{\|A \cup B\|}$	Surface-level token overlap between two texts	• Distinguishes lexical vs. structural robustness • Shows whether the model paraphrases or copies
Translation Similarity	Semantic similarity between translations of original vs. mutated text	• Shows whether the model captures intended contrasts or ignores structural cues
Entropy $\mathrm{H} = -\sum_t p(t) \, log \, p(t)$	Unpredictability or dispersion of the model's output distribution	• Identifies mutation types that increase uncertainty or introduce ambiguity
Mutation complexity $C_{mut} = \frac{EditDistance(x,x')}{max(\|x\|, \|x'\|)}$	How disruptive a transformation is relative to the original input	• Enables correlation between linguistic complexity and translation degradation • Maps competence across syntactic phenomena

Add/Delete characters- comparison

Add Character

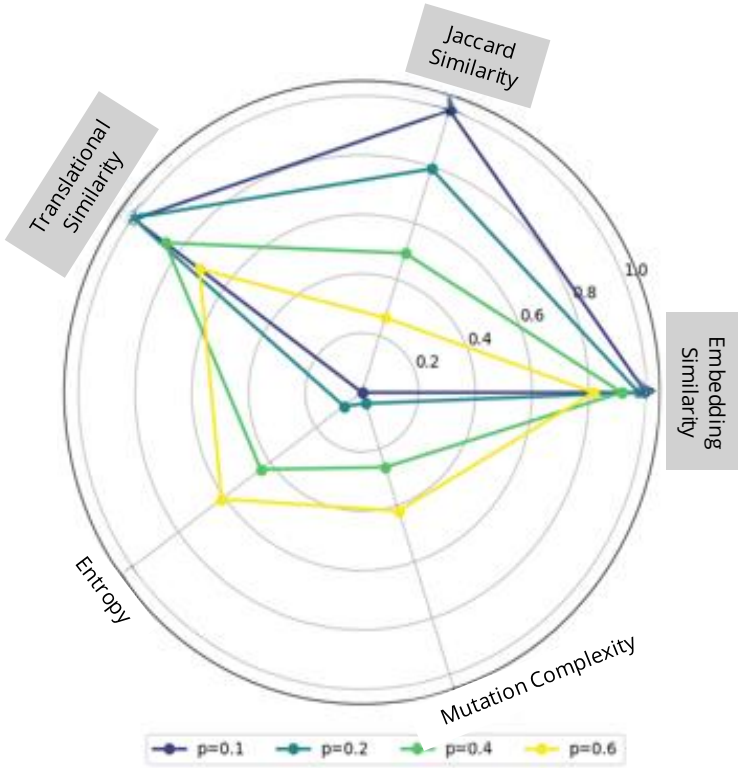


Delete Character

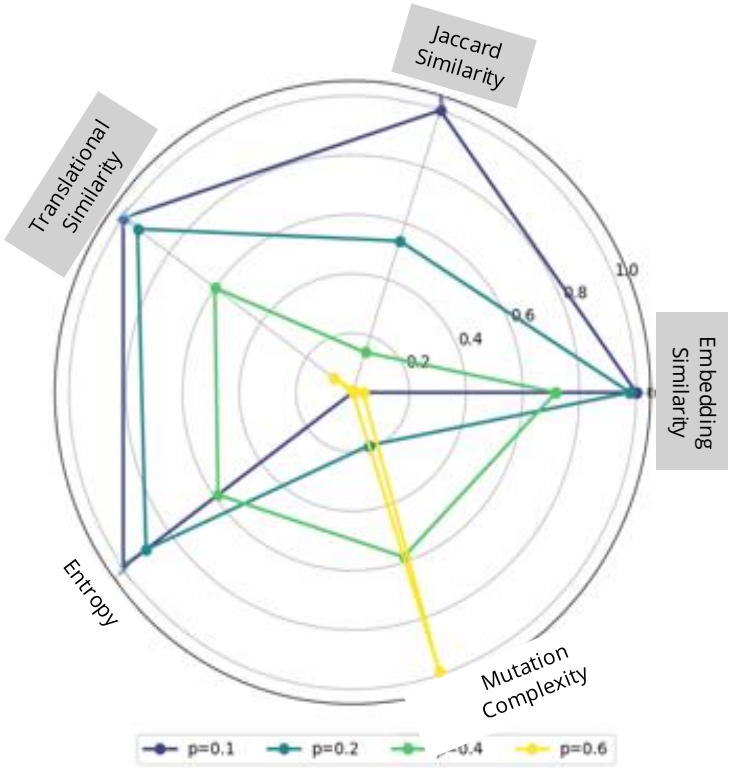


Add/Delete characters- comparison

Add Character



Delete Character



Mutation Probability

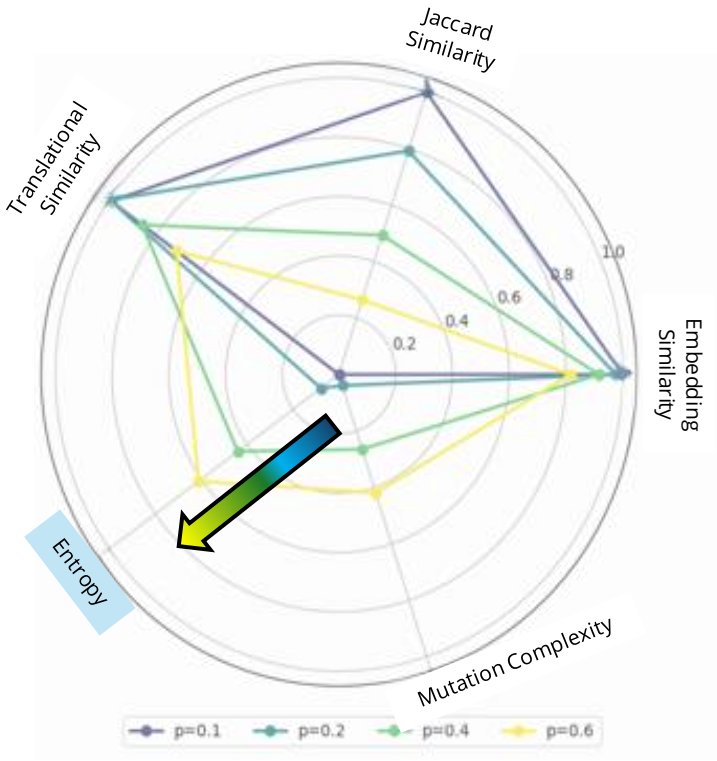
Similarity metrics

Higher Similarity metrics:

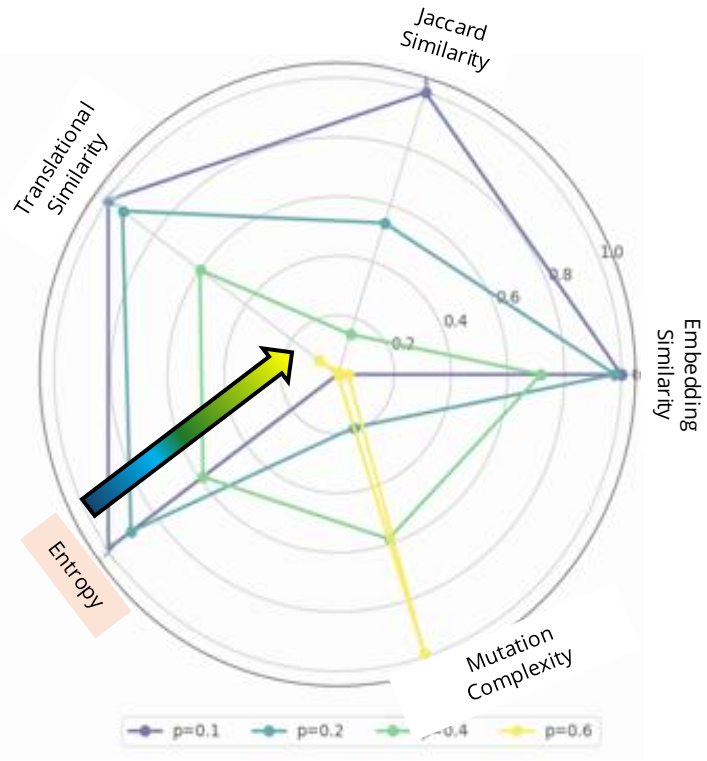
- Stronger semantic preservation
- More shared vocabulary
- Cross-lingual invariance

Add/Delete characters- comparison

Add Character



Delete Character



Entropy: *unpredictability of the output*

Higher value → more diverse responses

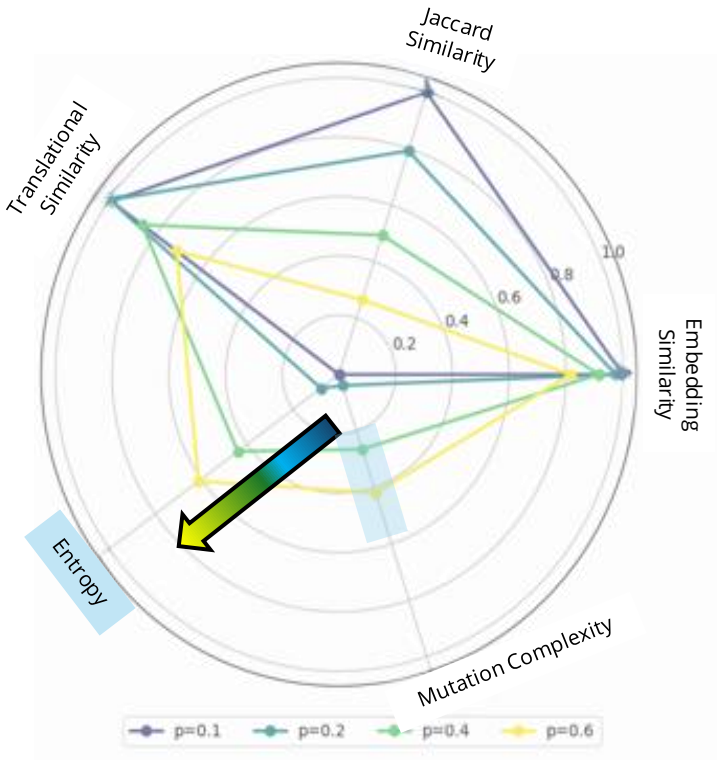
Lower value → deterministic behaviour

Increases for Add Char

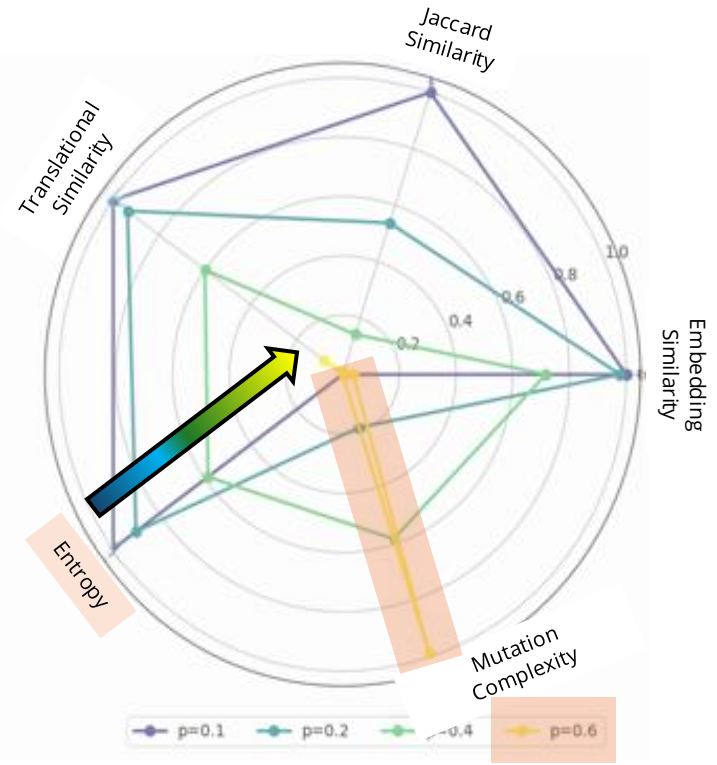
Decreases for Delete Char

Add/Delete characters- comparison

Add Character



Delete Character



Entropy: *Unpredictability of the output*

Higher value → more diverse responses

Lower value → deterministic behaviour

Increases for Add Char

Decreases for Delete Char

Mutation Complexity:

Disruption relative to original

Higher values → more perturbations

Max for Delete Char at 60% mutation

Deterministic

Translation unreliable!

Conclusion

1. **Semantic stability of LLMs,**
   - ✓ Under wide range of perturbations
   - ✓ Longer contexts → a strong buffer against noise

Conclusion

1. **Semantic stability of LLMs,**
   - ✓ Under wide range of perturbations
   - ✓ Longer contexts → a strong buffer against noise

2. **The vulnerability of character deletion highlights,**
   - ✓ Importance of structural cues
   - ✓ Linguistically principled frameworks

Conclusion

1. **Semantic stability of LLMs,**
   - ✓ Under wide range of perturbations
   - ✓ Longer contexts → a strong buffer against noise

2. **The vulnerability of character deletion highlights,**
   - ✓ Importance of structural cues
   - ✓ Linguistically principled frameworks

3. **Assessing LLM resilience with,**
   - ✓ Chomsky's Universal Grammar → deep structural invariants for mutation sets
   - ✓ Metamorphic testing → systematic manipulation of input text to check consistency

Demo of **QALITY**

With this theoretical background,
now Victor will talk about our agent **QALITY**.

Why does this matter for QAs

1. **Understanding your LLM helps you improve your prompts**
   - *Example:* Biases affects LLM the same way the output like POs poor requirements can end up with a solution that it's not what he wanted.
   - When is it not reliable. Breaking point.
   - Understanding that automatic translations can result in different branches of the user journey

2. **There's a whole new range of parameters to take into consideration when Automatizing QA tasks with LLM**

Contact us for further questions

Jalpa Soni
Senior Data Scientist at Capitole

jalpabensoni@capitole-consulting.com

Linked In: jalpa-soni

Victor Trujillo
QA Practice Lead at Capitole

victortrujillo@capitole-consulting.com

Linked In: victor-trujillo-fernandez-74388b136

expoQA®26

MADRID 26th, 27th & 28th May

Thank you for attending

expoqa.eu