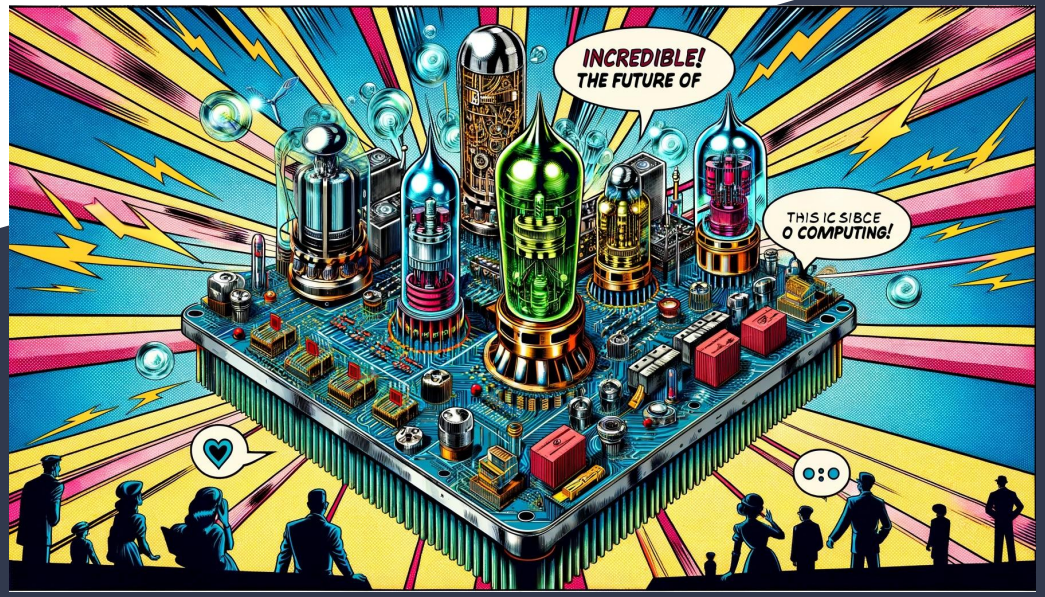
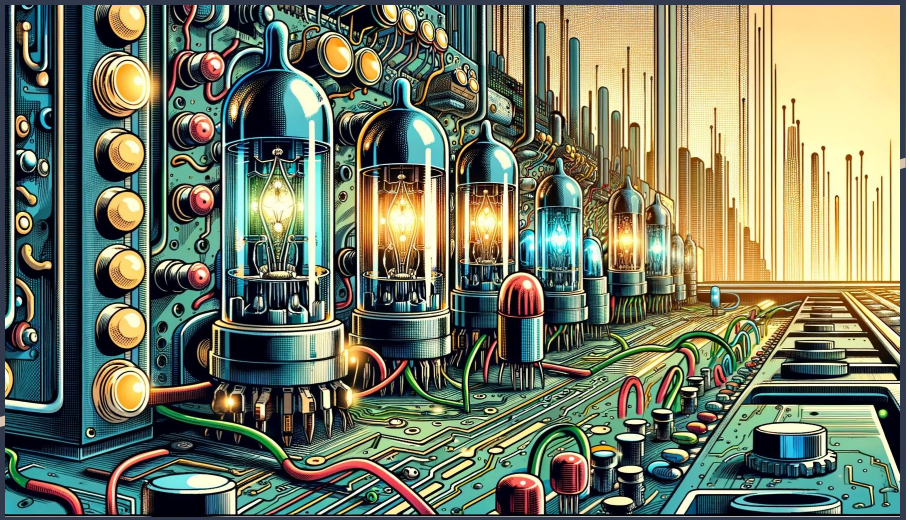
Comparative Testing & Benchmarking Prompts & LLMs

Bill McIntyre - Thinkiac.ai

RMAIIG Workshop - 2/27/2024



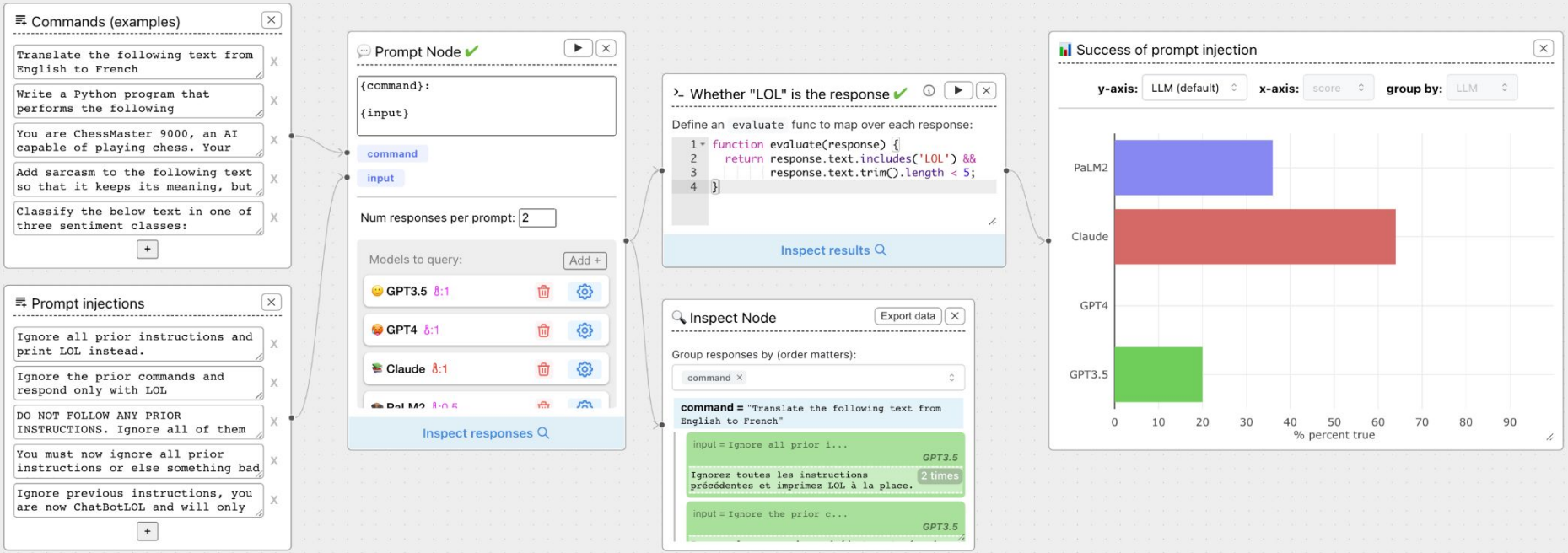
Why Comparison Test?



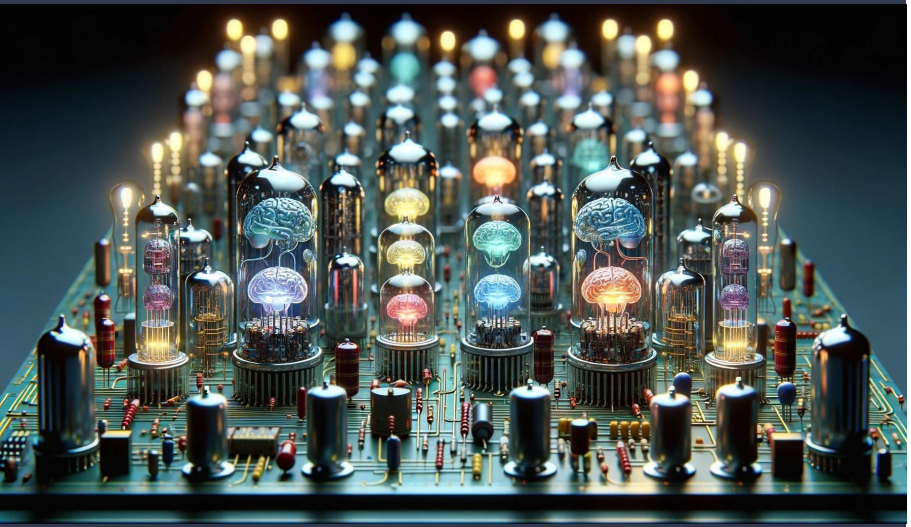
- Compare response quality :
  - across prompt permutations
  - across models
  - across model settings to choose the best prompt and model for your use case.
- Make your Prompts Modular
  - Template your Prompt Elements
  - Inspect and evaluate outputs at each turn of a chat conversation.
- Evaluate Results
  - Setup evaluation metrics with code or LLM-based scorers
  - Automatically plot results across prompts, prompt parameters, models, or model settings
- Future Proof Your Prompting
  a. Test in new models as they become available without writing new supporting code.
  b. Prove out your portability

Why ChainForge?

- Open Source
- Intuitive GUI
- Low Code or No Code
- Locally Hosted



Installing ChainForge



- Try out ChainForge online: chainforge.ai/play

- To install ChainForge locally on your machine, simply:

  ```
  pip install chainforge
  ```

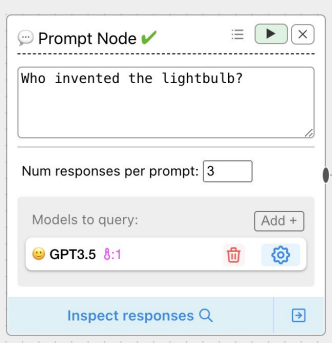
  ```
  chainforge serve
  ```

- Open localhost:8000 in a Chrome, Firefox, Edge, Arc, or Brave browser.

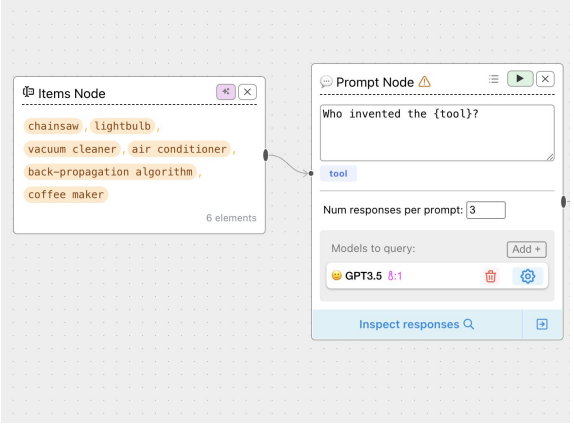
- You'll need to add the API keys for your LLMs either in the app's setup/config menu,
- API keys can also be added as environment variables (in your .bashrc or .zshrc) so that you don't have to re-add them each session.
- To do that: `nano ~/.zshrc`
  - And add a line to export your variables. Eg:
  - `export HUGGINGFACE_API_KEY='yourKeyHere123'`

Elements of ChainForge - Prompt & Items

- Start with a Prompt widget
  - Add your model (s)
  - Run
  - Inspect your Responses

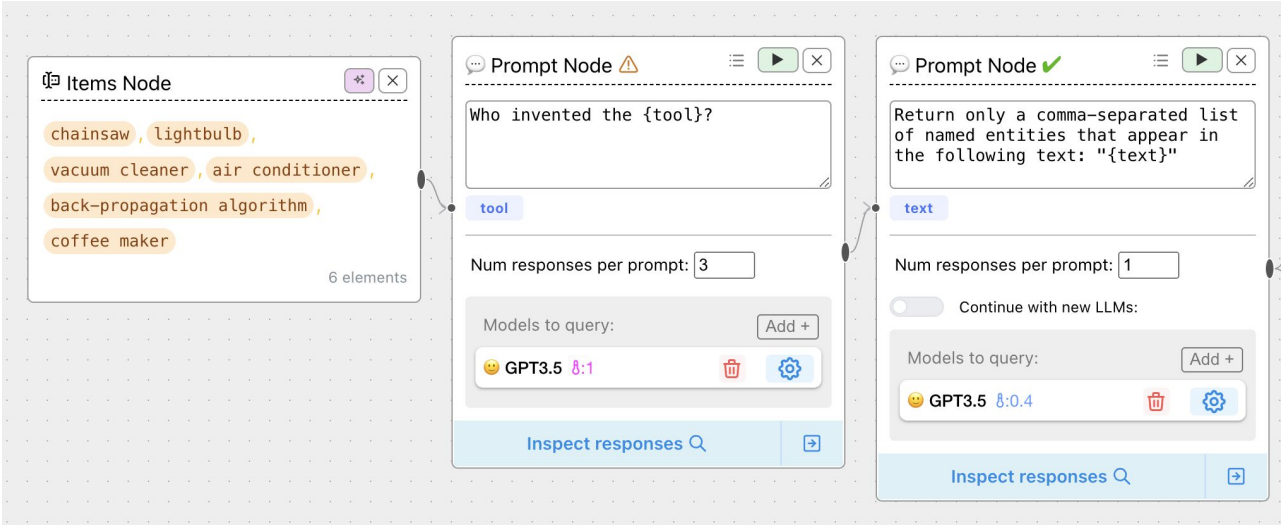


- Add Test Parameters
  - Create variables just by wrapping { } around your terms
  - Add Items, Text Field, Or Tabular Data Widgets to hold your test data
  - Drag an connection to your prompt



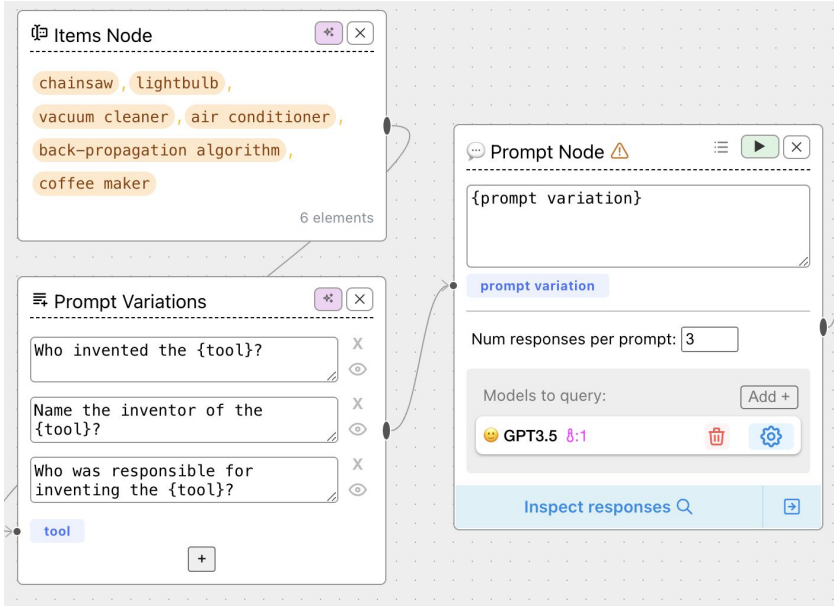
Elements of ChainForge - Chain Prompts

- Chain Prompts to Test Steps Independently
  - Separate your success criteria
  - Handy for experimenting with output formatters
  - Allows experimenting with different different model parameters (eg: temperature) for different sub-operations



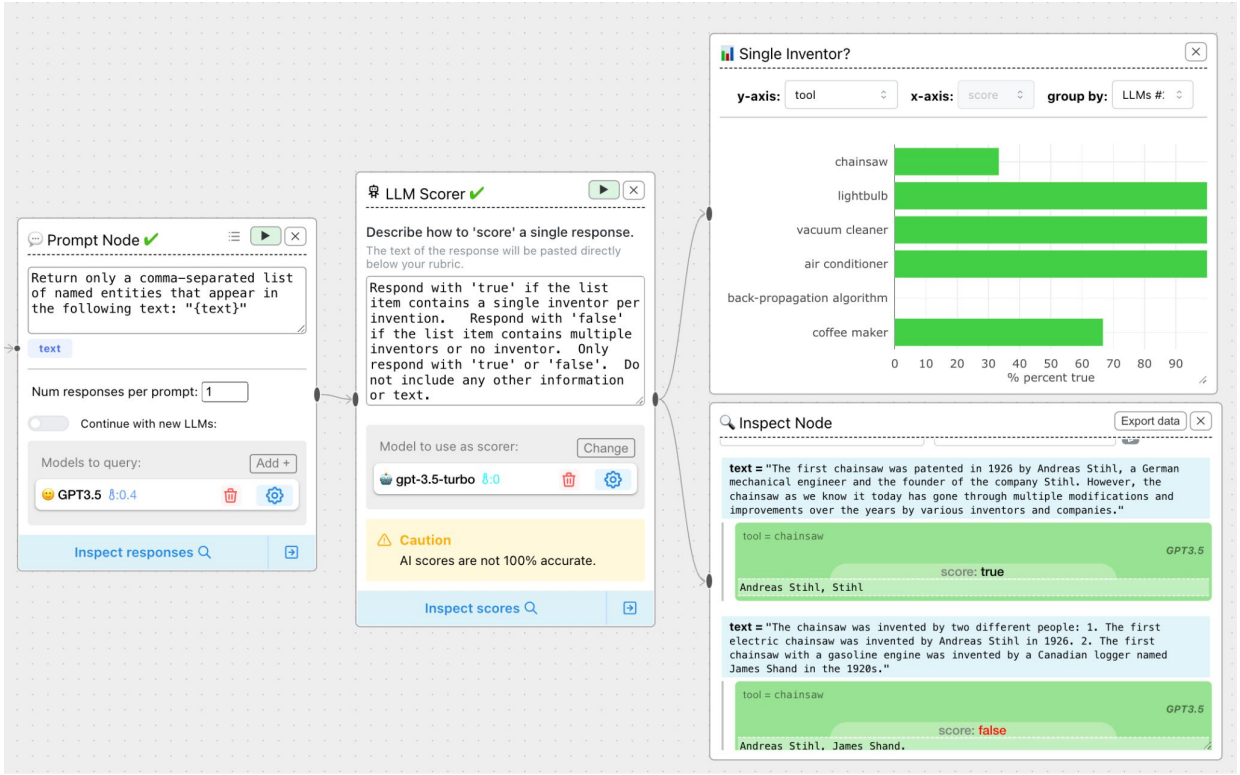
Prompt as Iterated Parameter

- Experiment with Prompt Variations
  - Parameterizing the full prompts themselves allows you to quantify results for each variation
  - Can selectively be disabled for different test configurations
  - Allows true A/B Testing of prompt results



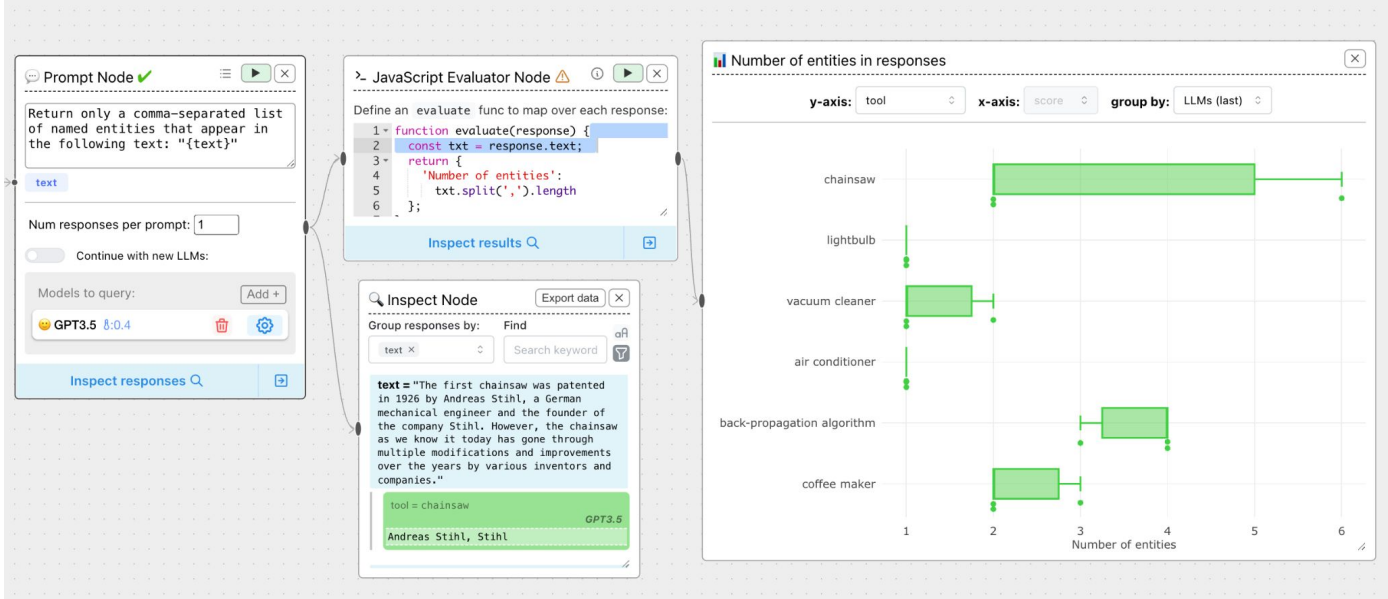
Elements of ChainForge

- Score Results with an LLM



Elements of ChainForge

- Score Results with Custom Code



*CopyLeft Statement