LLM Finetuning Research questions for Multi-Model Ensemble Training for Weird Machine Gadget Classification

1. Agreement Pattern Analysis

    1. Which examples cause disagreement between models?

All of them? They disagreed technically on all 50 examples that was validated. DistilGPT2 consistently predicts the correct Gadget Type while FLAN-T5 consistently suffers from repetition loops or hallucinations.

These example is an example of where a reasoning LLM would help, because FLAN-T5 made a minor typo but the gadgets were semantically the same.

Hypothesis to Test

1. Do disagreements correlate with excerpt length?

No, the disagreements appear systemic regardless of content, FLAN-T5 fails across every single example.

2. Are certain gadget types more ambiguous?

No, FLAN-T5 fails equally across all categories.

3. Does technical jargon cause confusion?

Yes, for FLAN-T5, likely triggering the repetition loops. FLAN-T5 often latches onto a piece of jargon and repeats it until it breaks down.

Example: FLAN-T5: Commutation-BID sys-initial sys-to-id-det-data-data sys-sys-inion a sys-inion sys-in-sys-data a sys-inion-data-as a sys-inion-i-diath a sys-inia-sys-inia-sys-inia-sys-inia-sys-inia-sys-in-ia-sys-inia-sys-inia-sys-inia-sys-inia-sys-in-ip-sys-sys-sys-sys-inia-sys-inia-sys-sys-inia-sys-inia-sys-inia-sys-sys

However, DistilGPT2 appears to be working fine and correctly identifying gadget types.

4. Do normalized types reveal that some "disagreements" are actually spelling variations?

Yes, but only for a small minority. In most cases normalization confirms that the disagreement is fundamental, rather than superficial.

Examples:
Example 4:Instruction: Identify weird machine TIMING/SYNCHRONIZATION gadgets in Logix 5000 (tasks, RPI,...

FLAN-T5: TIMing/Synchonization gadget

DistilGPT2: Timing/Synchronization gadget

Normalized FLAN-T5: timingsynchonization

Normalized DistilGPT2: timingsynchronization

Gold: gadget_type: Timing/Synchronization gadget; location: Produced/consumed tag RPIs...

Example 5:

Instruction: Identify weird machine SECURITY-PROCESSING gadgets in the excerpt and output gad...

FLAN-T5: SECURITY-Procsing gadget

DistilGPT2: Security-Processing gadget

Normalized FLAN-T5: securityprocsing

Normalized DistilGPT2: securityprocessing

Gold: gadget_type: Security-Processing gadget; location: Role and permission mapping l...

2. Architectural Comparison

Question: Does seq2seq (FLAN-T5) outperform causal LM (DistilGPT2)?

SUMMARY:

Total examples: 50

Full agreement: 0 (0.0%)

Disagreements: 50 (100.0%)

Format accuracy by model:

- flan-t5-small: 4.0%
- distilgpt2: 94.0%

No, casual LM outperformed seq2seq. With 50 examples validated by both models, FLAN-T5 had a format accuracy rate of 4% while DistilGPT2 had 94%. With DistilGPT2 consistently output short, valid class labels (e.g Control-Flow gadget). While FLAN-T5 faile dto adhere to the output format, suffering degenerate repetition (e.g., "A weird machine. A weird machine...").

Agreement with Gold Standard:

DistilGPT2 had a high agreement with the gold standard in the observed disagreements, where FLAN-T5 failed.