ANLP – EX 1

**Section 1 – Open Questions**

# QA datasets that annotate *intrinsic* concepts

* **QA-SRL 2.0**: Each predicate in a sentence is turned into question–answer pairs (“Who *bought* **what**?"”), giving full argument structure. The task probes *semantic role labeling*, a core property of compositional semantics independent of any downstream use.
* **QA-ZRE (Zero-Shot Relation Extraction)**: Facts such as *(Barack Obama, spouse, Michelle Obama)* are expressed as QA (“Who is Barack Obama married to?”), so models must recover binary relations directly from text—an intrinsic measure of relational semantics.
* **Quoref**: Questions require resolving intra-paragraph coreference (“What did *he* do after the game?”), so success hinges on the intrinsic ability to track entity mentions across clauses.

# Inference-time scaling methods

1. **Method comparison**
2. Self-Consistency:
3. **Single-GPU choice**

Given a single GPU with large memory and a *reasoning-heavy* scientific task, I would pick **Chain of Thought + Self-Consistency**:

* + It delivers strong logical performance with minimal engineering.
  + All *k* rationale samples can be generated in parallel batches on the same GPU, fully utilising the large memory bandwidth.
  + Unlike ToT or Reflection, it needs no stateful search tree and no read–write interaction with external tools, keeping VRAM usage predictable.
  + Empirically, accuracy saturates quickly with moderate *k* (8–16), so total inference time remains practical even on one device.