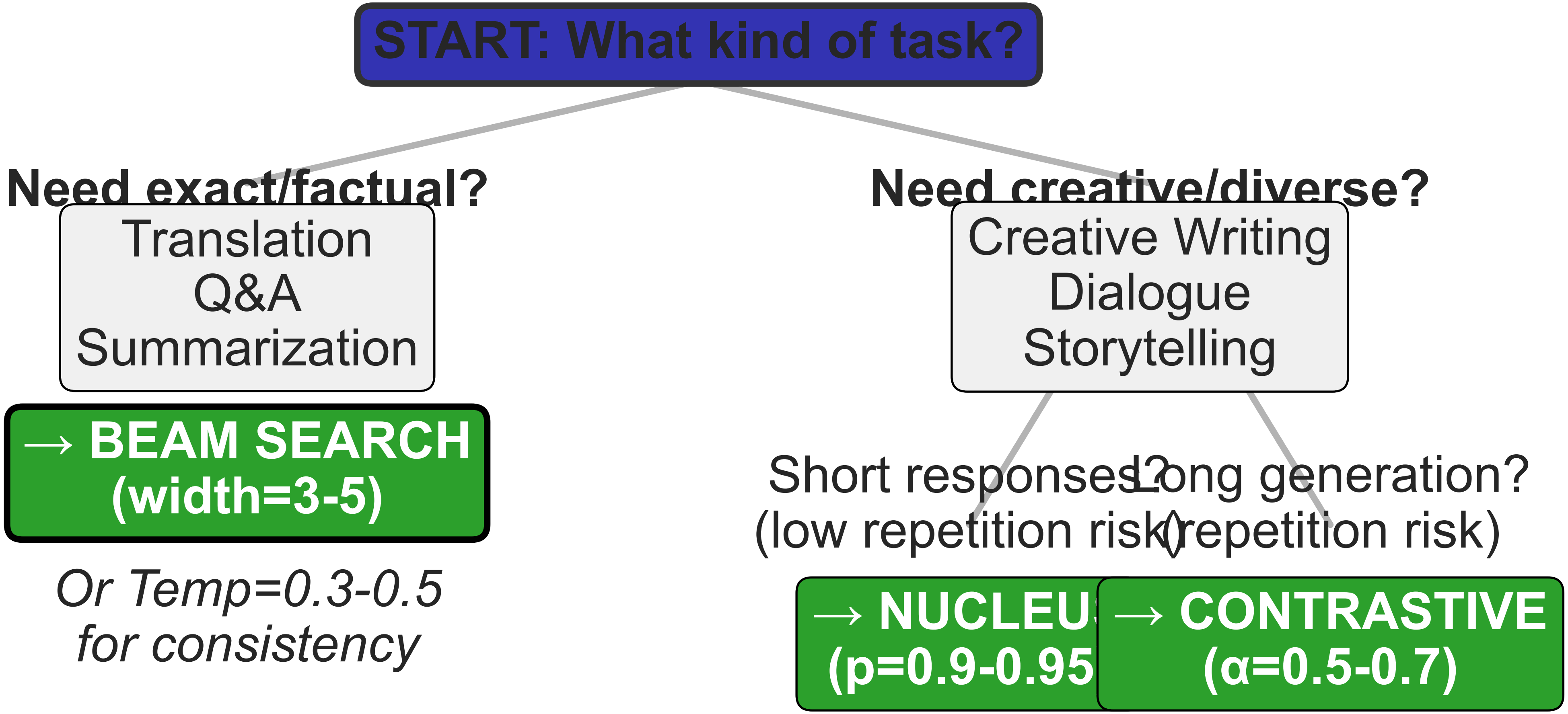
Choosing the Right Decoding Method



START: What kind of task?

Need exact/factual?

Translation
Q&A
Summarization

→ BEAM SEARCH
(width=3-5)

Or Temp=0.3-0.5
for consistency

Need creative/diverse?

Creative Writing
Dialogue
Storytelling

Short responses?
(low repetition risk)

→ NUCLEUS
(p=0.9-0.95

Long generation?
(repetition risk)

→ CONTRASTIVE
(α=0.5-0.7)

Special: Code Generation

→ Greedy or Beam (correctness critical)

→ *Then verify syntax/semantics*