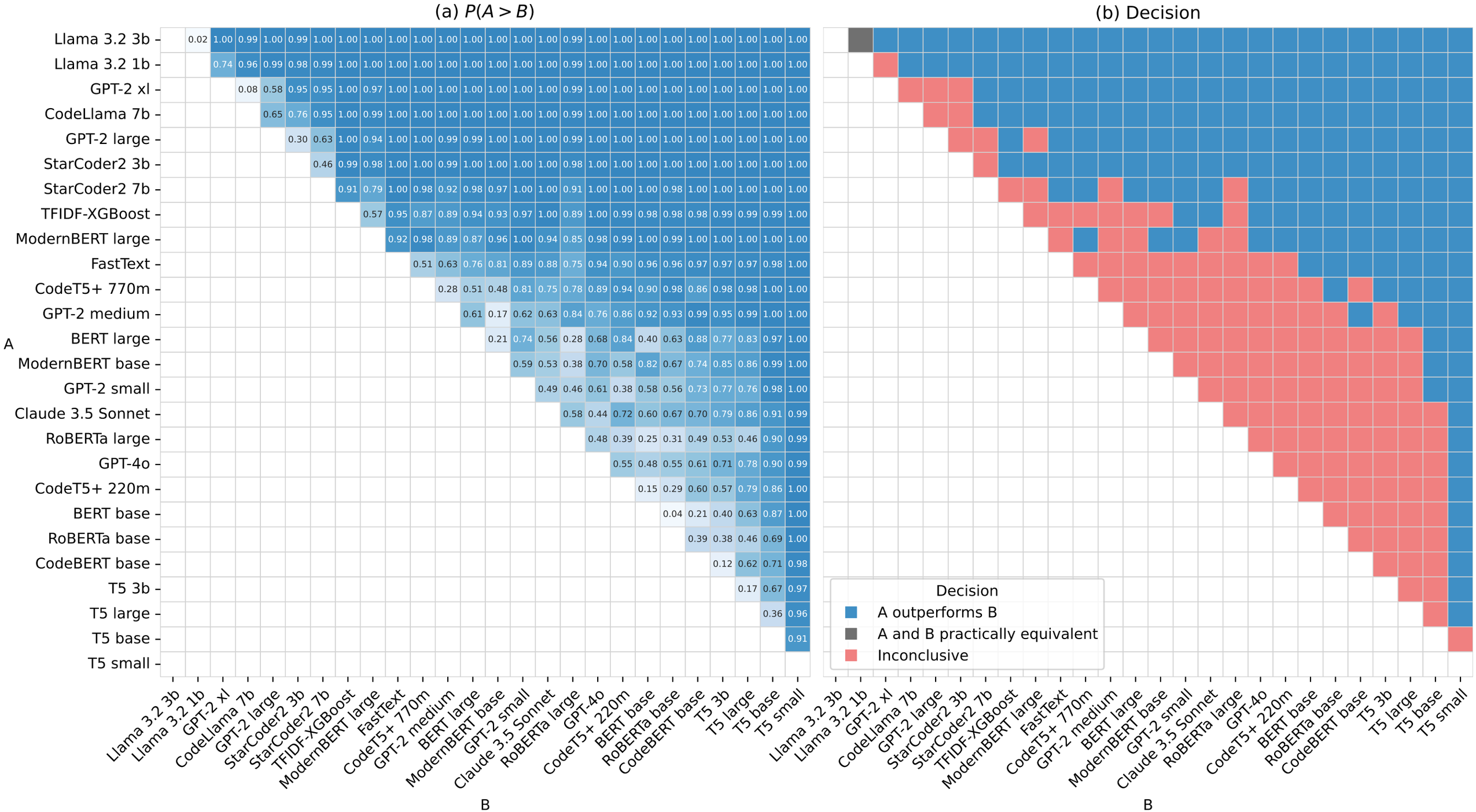
(a) $P(A > B)$
(b) Decision
A
B
Llama 3.2 3b
Llama 3.2 1b
GPT-2 xl
CodeLlama 7b
GPT-2 large
StarCoder2 3b
StarCoder2 7b
TFIDF-XGBoost
ModernBERT large
FastText
CodeT5+ 770m
GPT-2 medium
BERT large
ModernBERT base
GPT-2 small
Claude 3.5 Sonnet
RoBERTa large
GPT-4o
CodeT5+ 220m
BERT base
RoBERTa base
CodeBERT base
T5 3b
T5 large
T5 base
T5 small
Decision
A outperforms B
A and B practically equivalent
Inconclusive