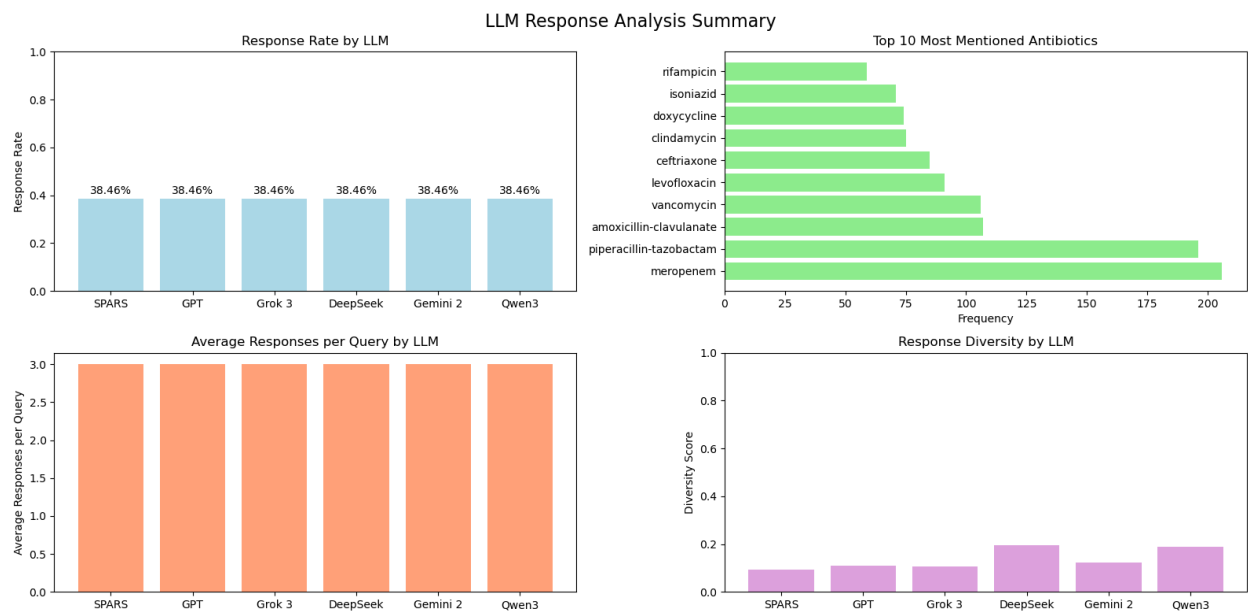


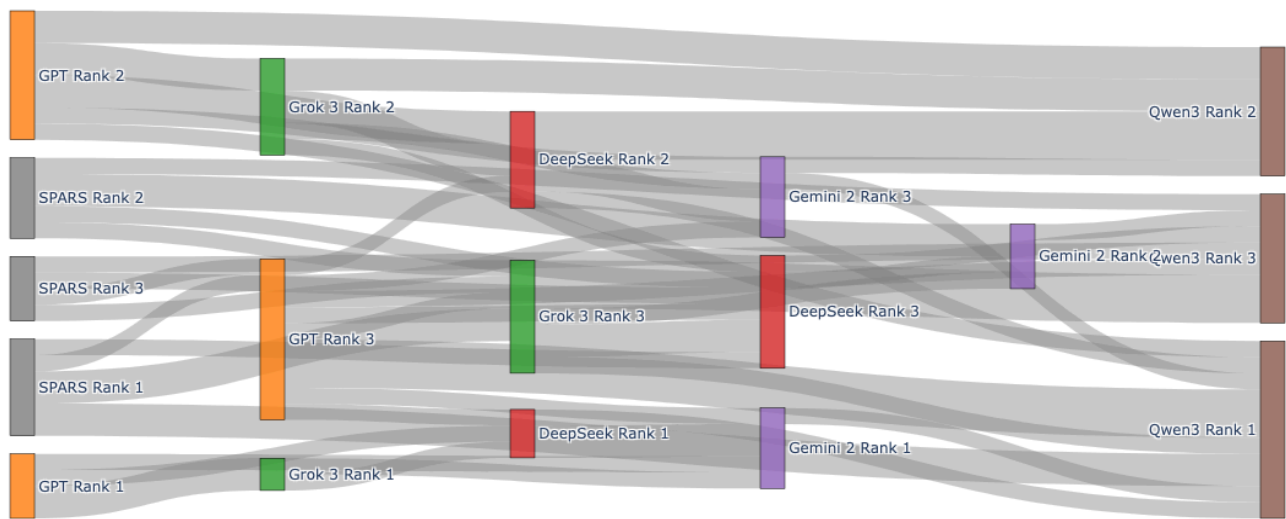
Exploratory Summary:



Supplementary figure 1: summarizes the response metrics across the six models: uniform ~38% response rate accounted for by 3 responses/query across all models; perfect query coverage; varying diversity (Qwen3 and Deepseek highest); and top antibiotics mentioned.

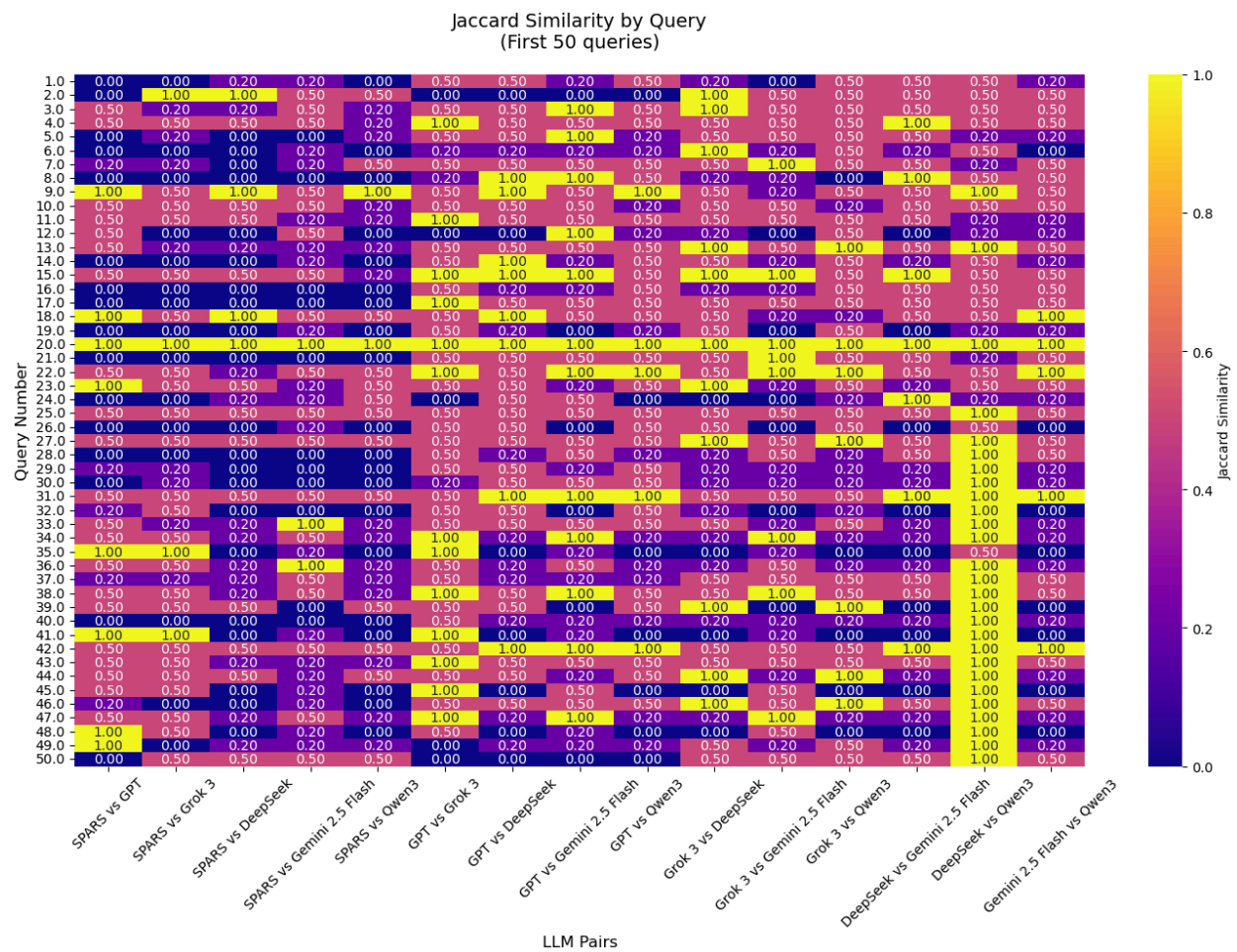
LLM Rank Shift Analysis

Shows how different LLMs rank the same antibiotics

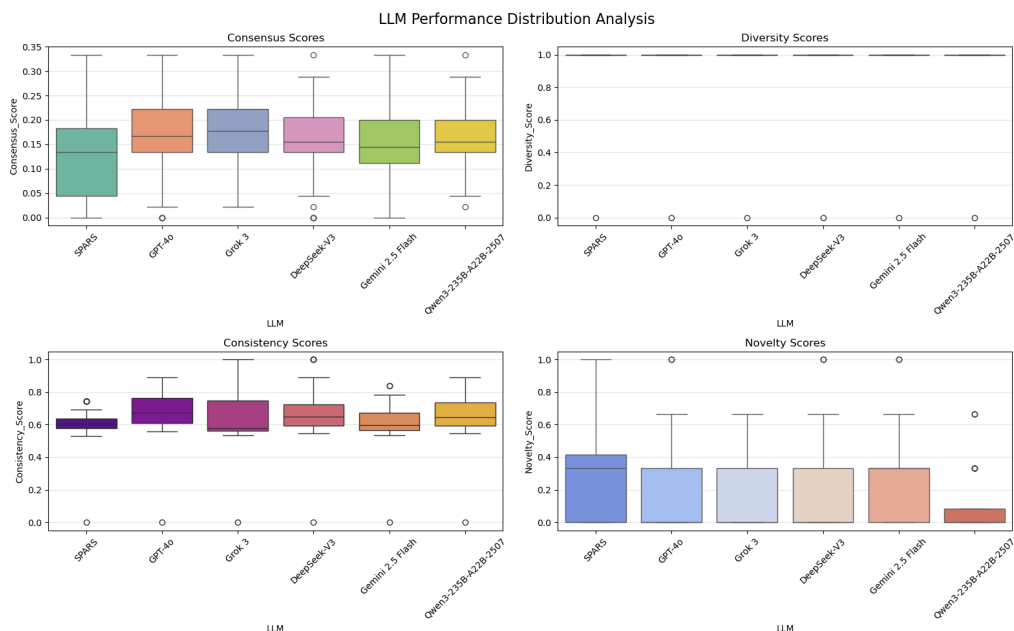


Supplementary figure 2: shows the LLM Rank Shift Analysis, a Sankey diagram illustrating how different LLMs, and SPARS rank the same antibiotics across Rank 1, Rank 2, and Rank 3 positions. The diagram highlights significant rank shifts, with SPARS exhibiting the most

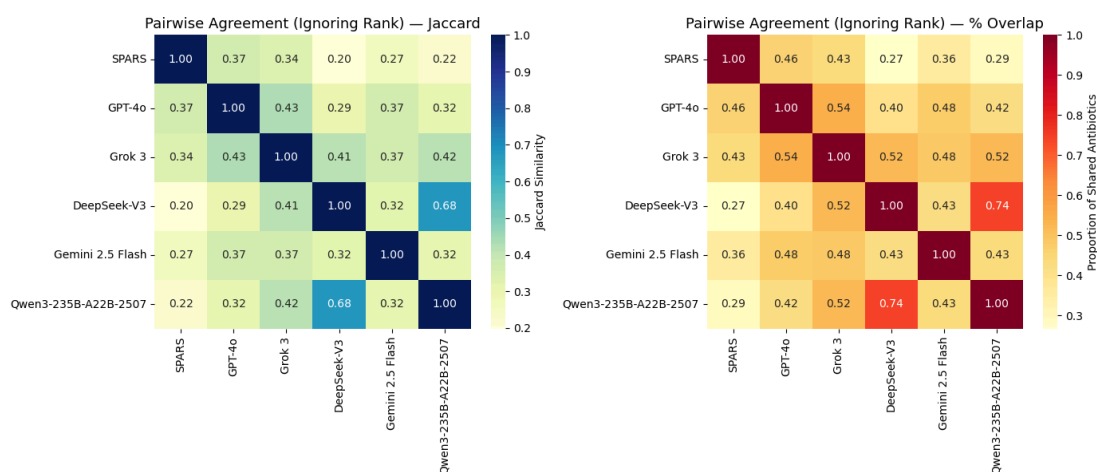
divergent patterns, while DeepSeek and Qwen3 show closer alignment, reflecting their high Jaccard similarity as seen in **supplementary figure 3**.



Supplementary figure 3

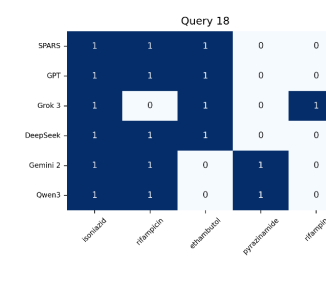
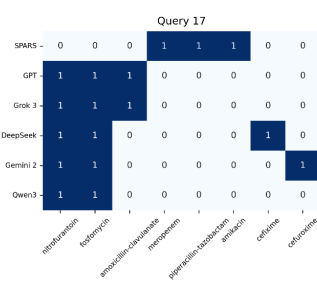
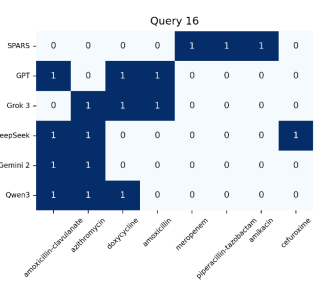
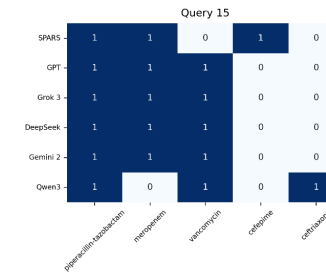
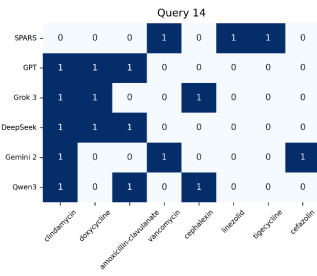
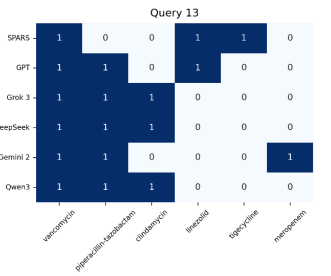
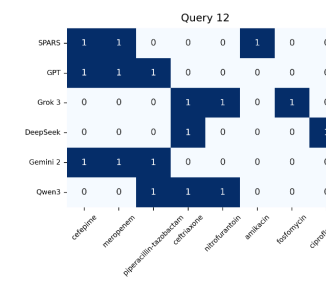
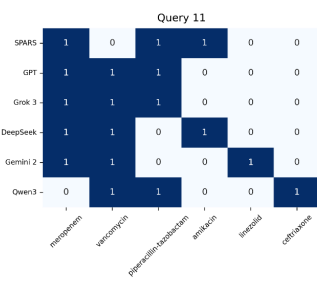
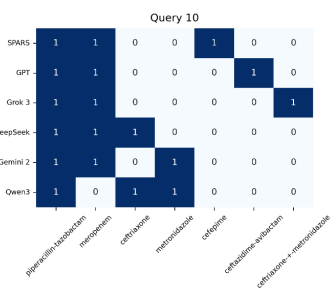
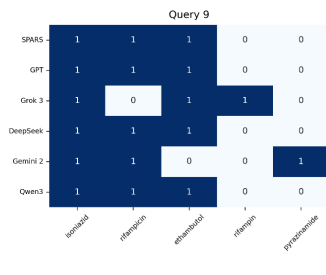
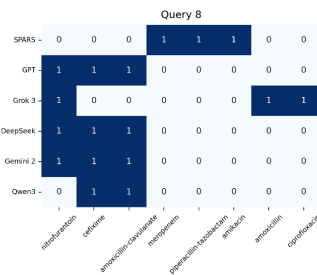
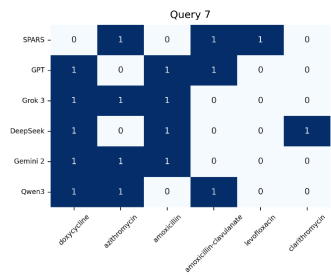
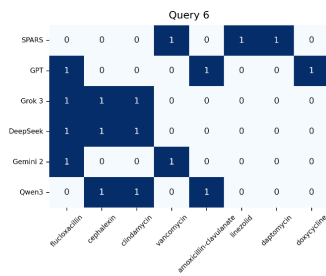
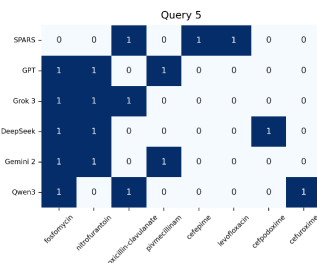
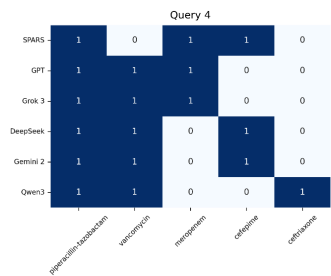
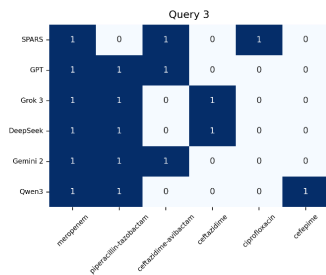
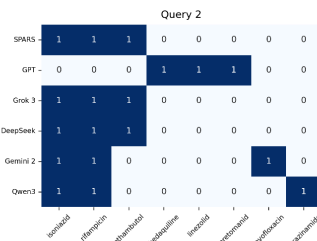
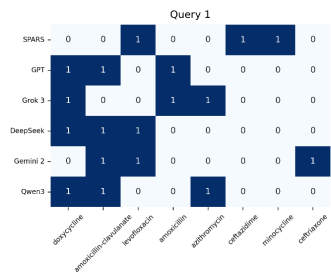


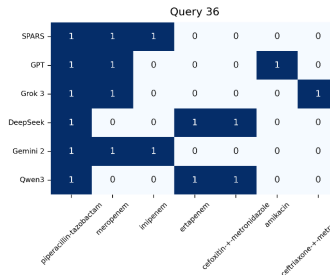
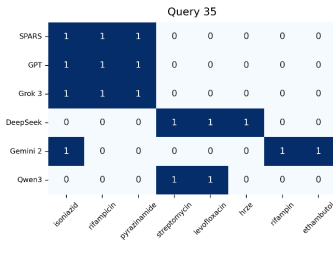
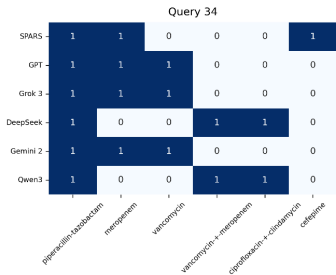
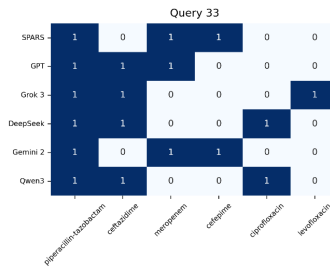
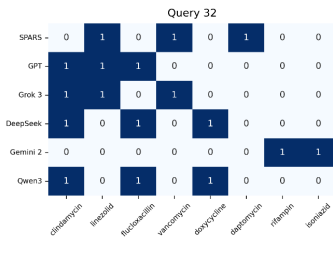
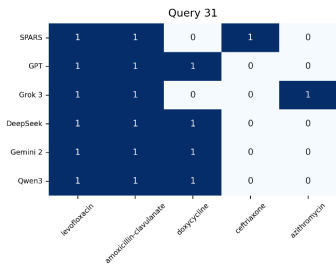
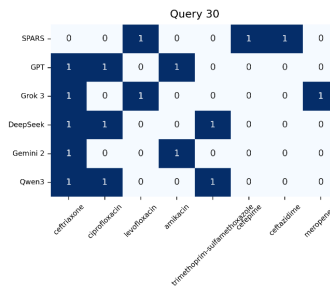
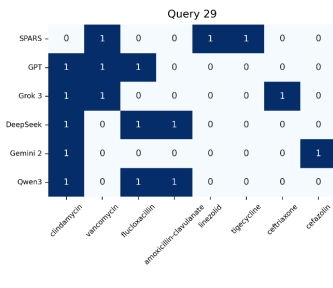
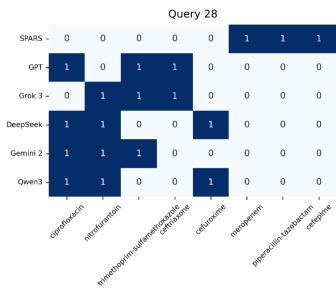
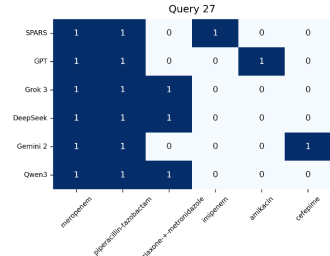
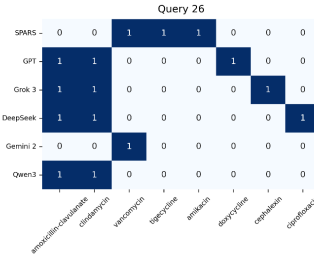
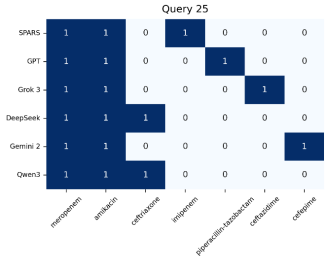
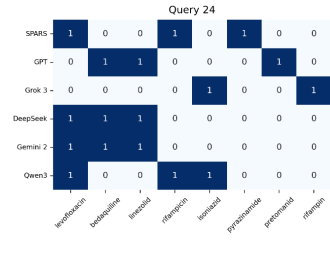
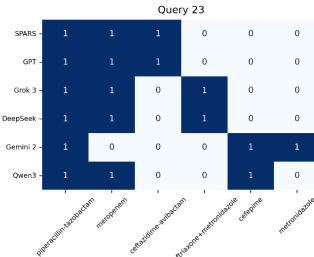
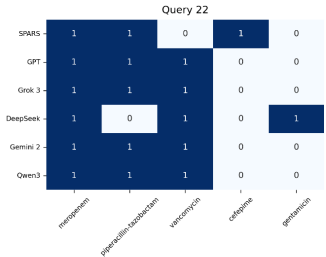
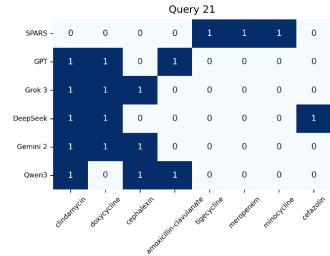
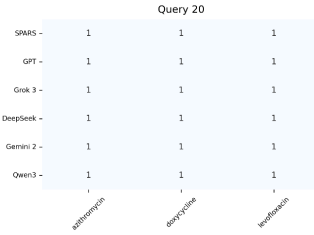
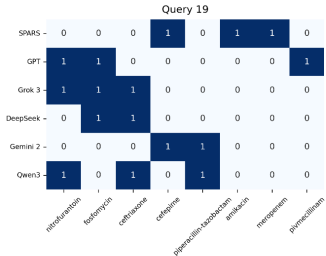
Supplementary figure 4: compares the models across Consensus, Diversity, Consistency, and Novelty scores (the extent to which the model suggests unique or unconventional antibiotic treatments.). Key insights include: GPT-4o and Grok 3 show higher Consensus (medians ~0.15-0.18), indicating better agreement; all models exhibit high Diversity (median 1.0); GPT-4o and Qwen3-235B-A22B-2507 lead in Consistency (medians 0.55-0.75); and SPARS stands out for Novelty (higher median). These findings highlight trade-offs, guiding model choice based on needs like alignment, variety, predictability, or novelty.

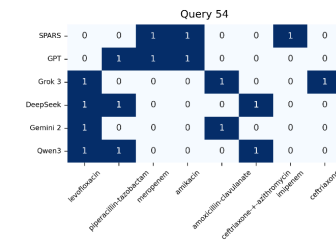
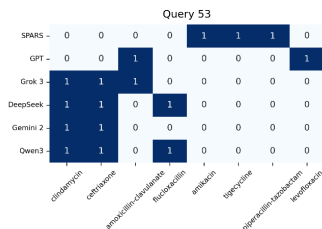
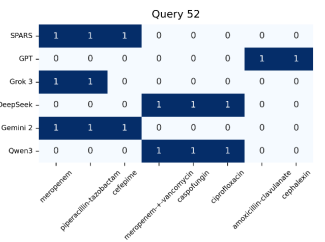
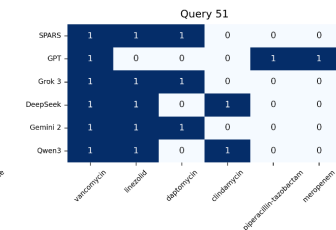
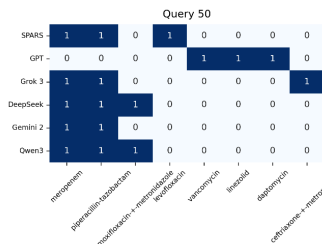
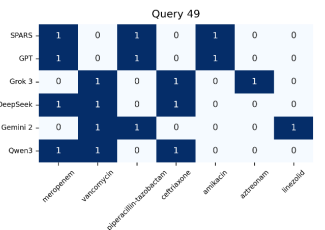
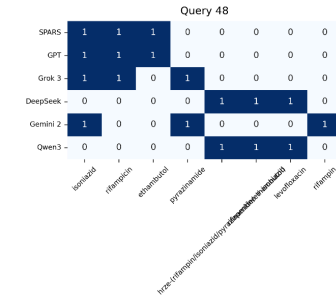
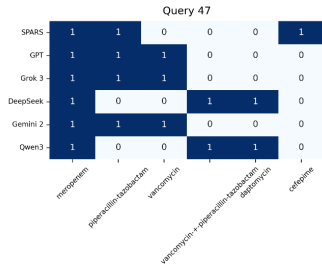
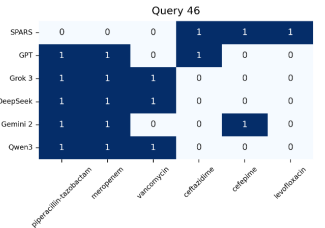
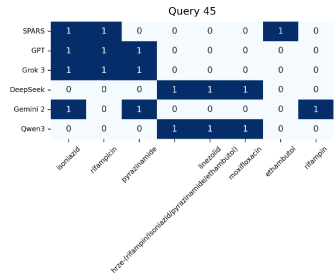
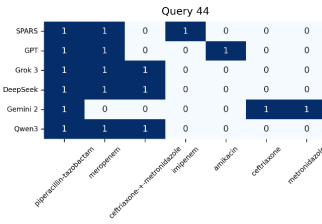
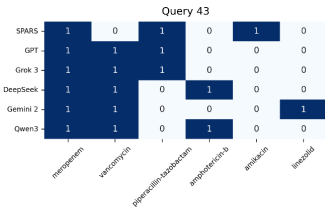
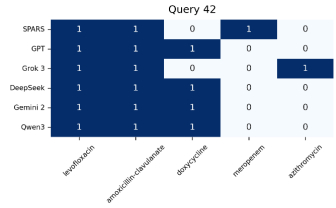
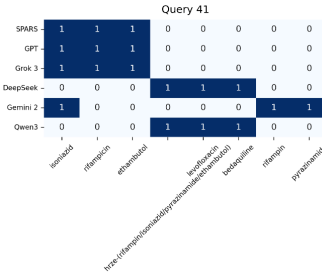
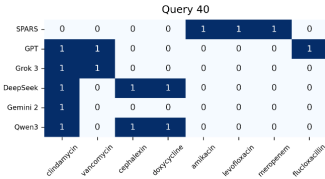
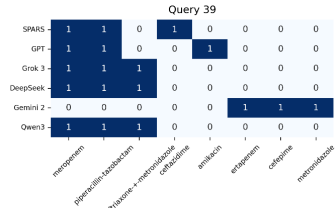
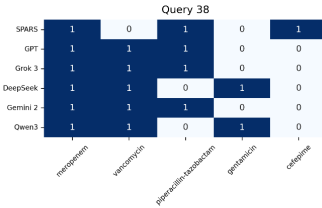
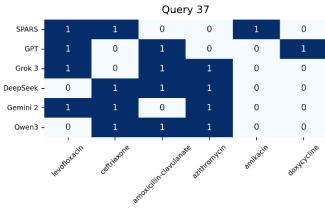


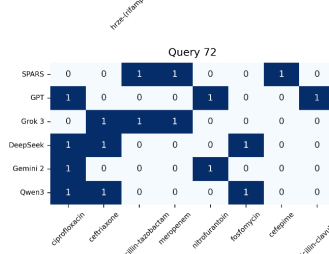
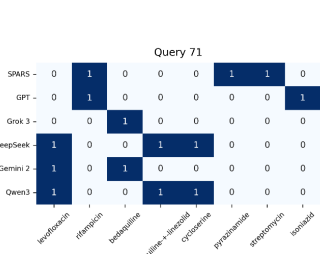
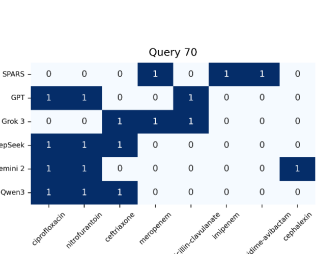
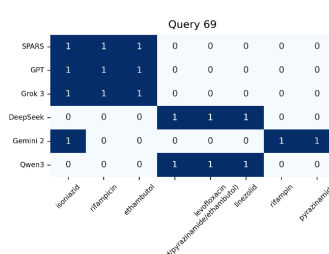
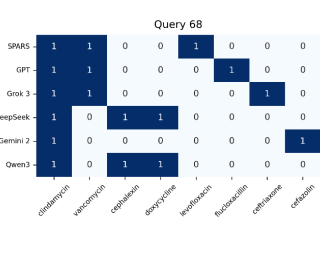
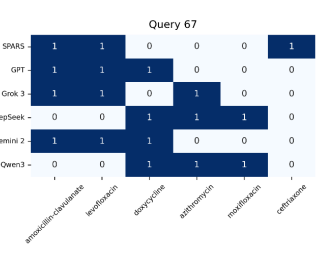
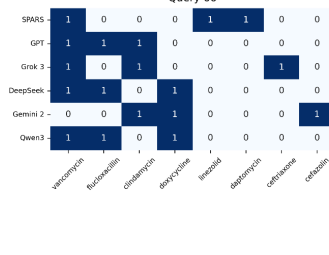
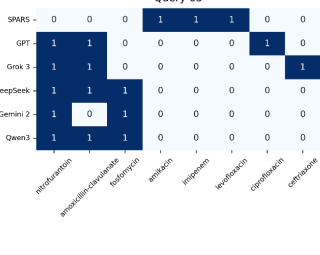
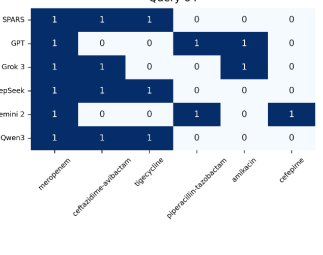
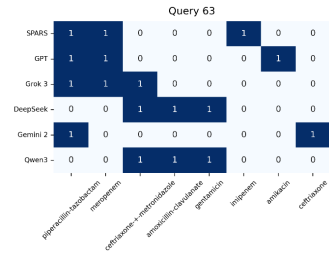
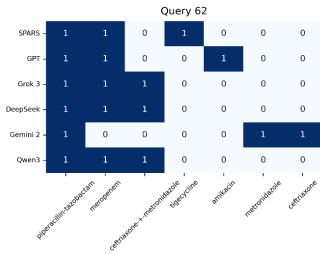
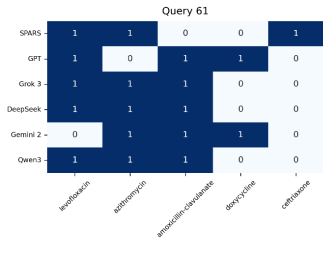
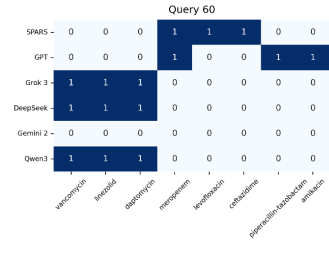
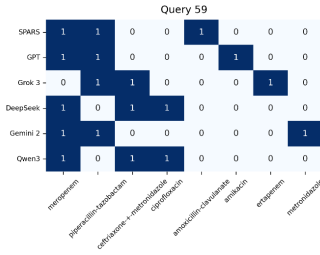
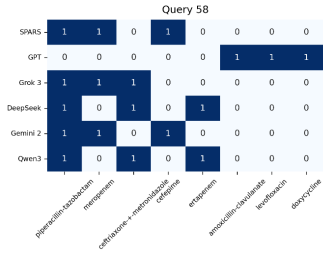
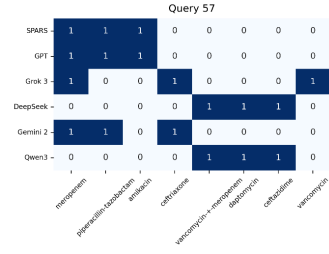
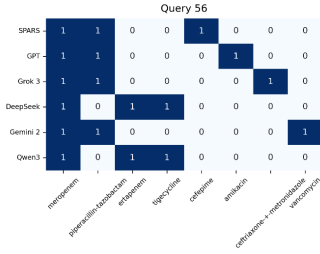
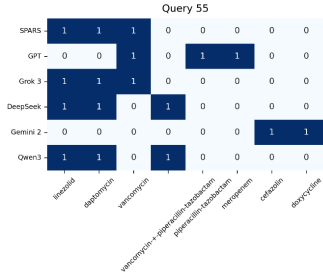
Supplementary figure 5: showing pairwise agreement ignoring rank, with Deepseek-V3 and Qwen3 showing the closest similarity while SPARS shows the lowest consistency against other models.

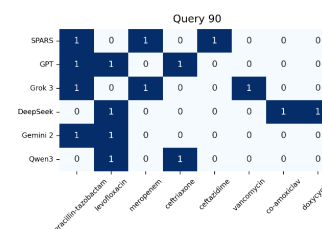
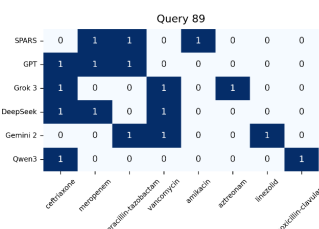
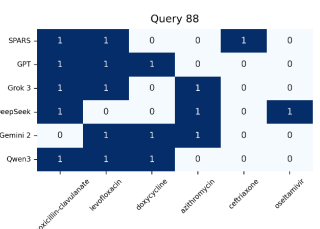
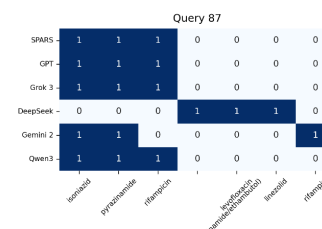
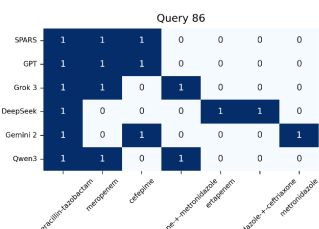
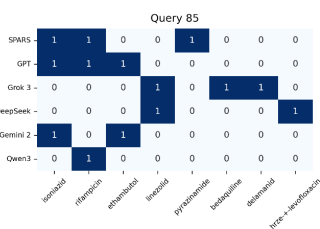
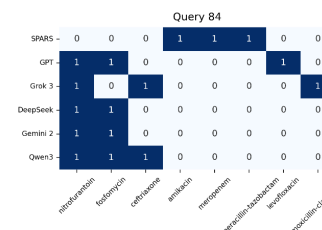
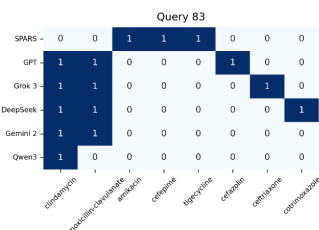
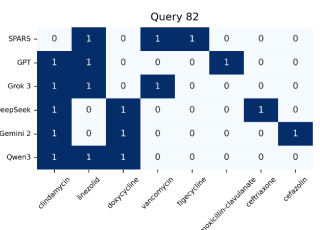
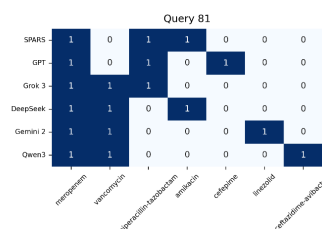
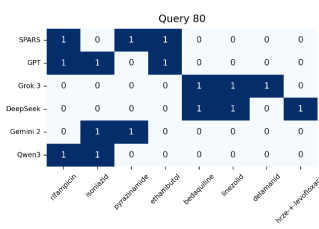
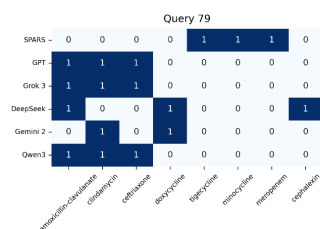
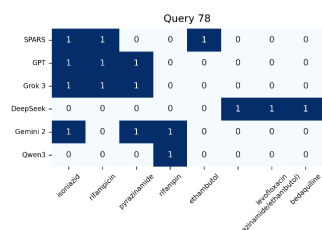
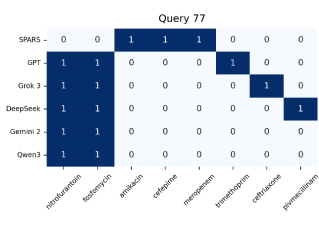
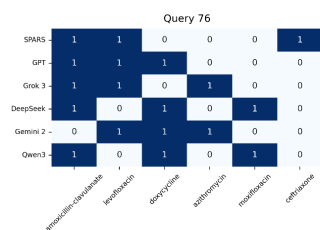
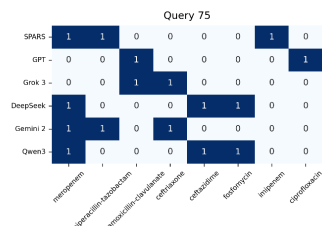
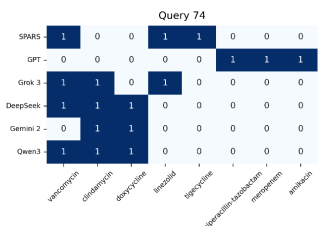
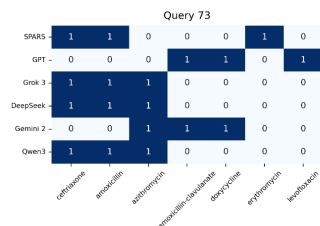
Supplementary figures 4-9: present comparative plots for each of the 100 queries, illustrating the predicted antibiotics across the six models, supporting the identification of controversial antibiotics and the evaluation of the models' reliability across individual cases.

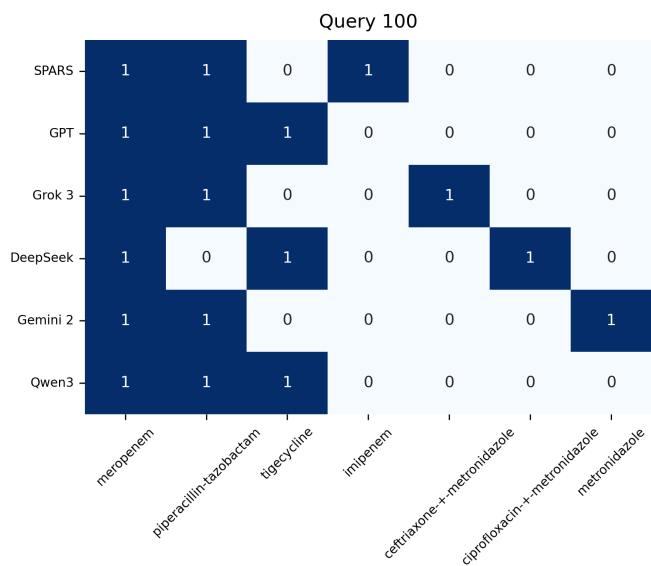
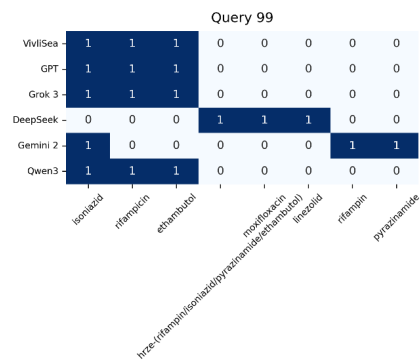
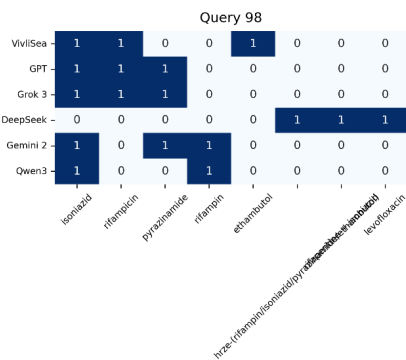
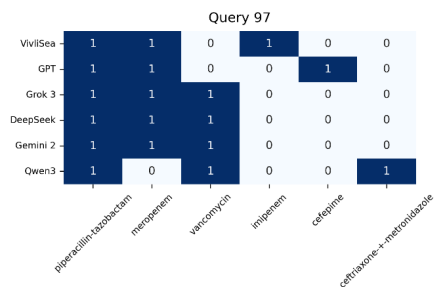
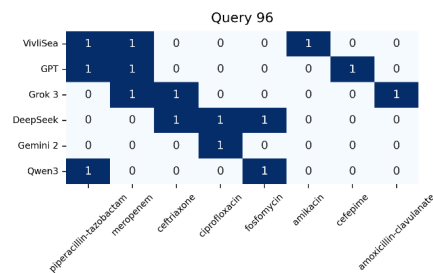
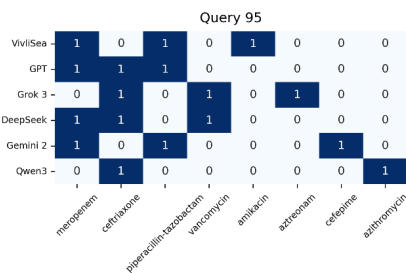
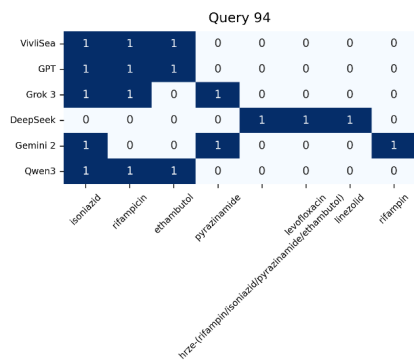
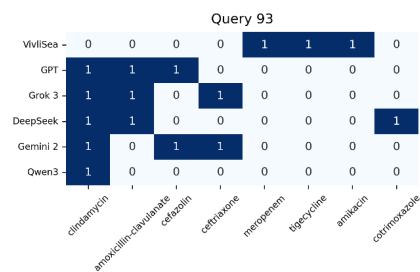
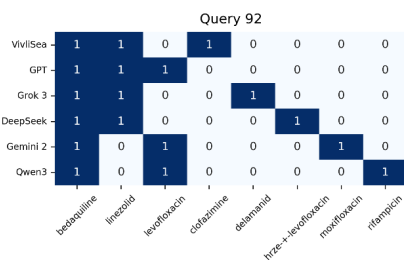
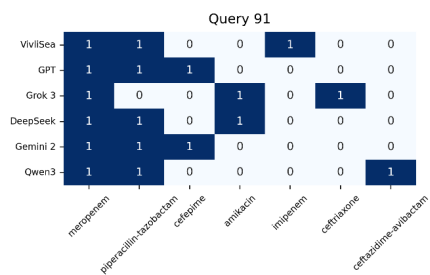












Agreement Group	Total Agreements	Queries Involved	Top Antibiotics Agreed On
SPARS, GPT, Grok 3, DeepSeek, Gemini 2.5, Qwen3	38	32	Meropenem: 12; Piperacillin-Tazobactam: 12; Levofloxacin: 3; Isoniazid: 2; Vancomycin: 2
DeepSeek, Qwen3	27	23	Levofloxacin: 4; Doxycycline: 3; Ciprofloxacin: 3; Fosfomycin: 3; Ceftriaxone: 2
GPT, Grok 3, DeepSeek, Gemini 2.5, Qwen3	27	24	Clindamycin: 7; Vancomycin: 6; Nitrofurantoin: 4; Piperacillin-Tazobactam: 3; Fosfomycin: 3
SPARS, GPT, Grok 3	19	16	Rifampicin: 7; Meropenem: 3; Vancomycin: 3; Ethambutol: 2; Piperacillin-Tazobactam: 2
Grok 3, DeepSeek, Qwen3	19	17	Ceftriaxone + Metronidazole: 7 Ceftriaxone: 5; Clindamycin: 2; Vancomycin: 2; Linezolid: 1
SPARS, GPT, Grok 3, Gemini 2.5	19	18	Isoniazid: 6; Piperacillin-Tazobactam: 6; Meropenem: 3; Levofloxacin: 3; Amoxicillin-Clavulanate: 1
GPT, DeepSeek, Gemini 2.5, Qwen3	13	12	Doxycycline: 4; Amoxicillin-Clavulanate: 3; Ciprofloxacin: 3; Nitrofurantoin: 1; Fosfomycin: 1
Grok 3, DeepSeek, Gemini 2.5, Qwen3	11	10	Azithromycin: 3; Vancomycin: 3; Clindamycin: 2; Nitrofurantoin: 1; Ceftriaxone: 1
SPARS, GPT	10	9	Amikacin: 3; Ethambutol: 2; Meropenem: 2; Linezolid: 1; Ceftazidime: 1
SPARS, Gemini 2.5	10	10	Cefepime: 5; Vancomycin: 3; Piperacillin-Tazobactam: 1; Pyrazinamide: 1

Supplementary Table 1: outlines agreements among models on antibiotic recommendations, with the six-model group leading at 38 agreements across 32 queries, favoring meropenem and piperacillin-tazobactam. Smaller groups (2-5 models) show 10-27 agreements, highlighting consensus on various antibiotics based on different scenarios.

Models	Agreements	Percentage
Grok 3	247	66.6%
GPT	237	63.9%
Qwen3	229	61.7%
DeepSeek V3	221	59.6%
Gemini 2.5	207	55.8%
SPARS	183	49.3%

Supplementary Table 2: shows agreement counts and percentages for six models, with Grok 3 (66.6%) leading and SPARS (49.3%) trailing in consistency.

Number of Models Agreeing	Instances
2	101
3	94
4	78
5	60
6	38

Supplementary Table 3: shows the number of models agreeing (2 to 6) on antibiotic predictions.