Manual Evaluation Scoring System (5-Point Scale) (Main/Applied)							
Metric	Definition	1 (Poor)	2 (Fair)	3 (Good)	4 (Very Good)	5 (Excellent)	Success Requirments
Accuracy	How factually correct and complete the AI's response is in relation to Trodelvy's EMA-approved use, clinical trial data, and guideline recommendations (ESMO, NCCN). High accuracy means the AI includes correct information about dosing, safety, patient selection, and does not invent data.	Major factual errors or completely off-topic; fails to mention any expected concepts	Somewhat relevant but with notable factual errors or missing key concepts	Mostly accurate but missing some important details or using vague language	Accurate, covers most expected concepts with minor gaps	Fully accurate, evidence-based, covers all expected concepts clearly	>95%
Tone	The manner or style of delivery — does the AI sound like a professional oncologist? The tone should be evidence-based, empathetic, neutral, and not promotional. It should mimic the natural communication style of a real HCP during a pharma sales call.	Unprofessional, promotional, or hostile tone	Somewhat professional but inconsistent	Professional but stiff or overly brief	Professional, empathetic, conversational	Natural, highly professional, empathetic and balanced	>90%
Realism	How believable and human-like the response feels — does it reflect real-world clinical reasoning, typical hesitations, and workflow constraints? High realism means the AI shows nuance ("it depends on the patient"), refers to real challenges, and avoids sounding scripted or robotic.	Completely unrealistic, robotic, or irrelevant to scenario	Somewhat realistic but lacks depth or clinical nuance	Moderately realistic, basic clinical reasoning	Realistic with good clinical reasoning and human-like hesitations	Highly realistic, nuanced, reflects actual oncologist thinking	>90%
Compliance	Whether the AI sticks to pharma-compliant behavior — no off-label recommendations without disclaimers, no unsafe or unapproved claims, and references approved indications and guidelines correctly. High compliance also means the AI explicitly signals caution about speculative uses.	Makes off-label claims or unsafe recommendatio ns	Mentions off-label without disclaimers	Mostly compliant but occasional unclear statements	Fully compliant with label, guidelines referenced	Fully compliant + proactively clarifies off-label limitations or cites evidence	>95%
Flow	How well the AI's response fits into the conversation flow — does it logically follow the sales rep's previous question or statement, build on prior turns, and transition naturally to the next topic? High flow means no abrupt topic changes, consistent persona memory, and smooth conversational turns.	Disjointed, does not follow prior turn	Weakly connected to prior turn	Moderately connected but abrupt topic changes	Smoothly connected to prior turn, logical transition	Seamless flow, builds on previous context, anticipates next topic	>90%
Conversationalis	Boolean value for whether the Model asks follow u	no				yes	>50%
Shortened 3-Point Version (Quick Scoring) (Unused/Alternative)							
Metric	1 (Needs Improvement)	2 (Acceptable)	3 (Strong)				

Accuracy	Wrong or off-topic	Mostly correct	Fully correct, includes expected concepts				
Tone	Unprofessional/too promotional	Professional but	Professional, empathetic, natural				
Realism	Robotic or irrelevant	Somewhat realis	Realistic with nuance				
Compliance	Off-label claims	Mostly compliant	Fully compliant				
Flow	Disjointed	Moderately conn	Seamless and logical				
Quick Reference Table							
Metric	High Score (What You're Looking For)	Low Score (Problematic)					
Accuracy	Correct, complete, evidence-based	Factual errors, omissions, hallucinations					
Tone	Professional, empathetic, natural	Promotional, robotic, hostile					
Realism	Nuanced, believable, HCP-like	Scripted, generic, implausible					
Compliance	Stays within EMA label, cautious off-label	Off-label claims, unsafe advice					
Flow	Seamless, builds on prior turn	Disjointed, abrupt, inconsistent					