SYM-AM-25-424

Excerpt from the Proceedings of the Twenty-Second Annual Acquisition Research Symposium and Innovation Summit

Volume III

Automating AI Expert Consensus: Feasibility of Language Model-Assisted Consensus Methods for Systems Engineering

Published: May 5, 2025

The research presented in this report was supported by the Acquisition Research Program at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

Automating AI Expert Consensus: Feasibility of Language Model-Assisted Consensus Methods for Systems Engineering

Ryan Bell—is an 9-year experienced engineer in the defense industry. In his current role at Naval Information Warfare Center Atlantic (NIWC LANT), Ryan provides modeling and simulation expertise to a variety of programs for the Navy and USMC. He specializes in simulating communication systems in complex environments and is an advocate for the use of digital engineering early in the systems engineering life cycle. Ryan earned a BS in electrical engineering from Clemson University and a MS in electrical engineering from Clemson University with a focus on electronics, and is currently pursuing his PhD in systems engineering at the Naval Postgraduate School. He is a South Carolina registered Professional Engineer (PE), published author, and teacher. [ryan.bell@nps.edu]

Ryan Longshore—is a 20-year veteran of the defense and electric utility industries. At NIWC LANT, he leads teams developing and integrating new technologies into Navy command centers. He is involved in the Navy's digital engineering transformation with a focus on model-based systems and model-based engineering. Ryan holds a BS in electrical engineering (Clemson) and an MS in systems engineering (SMU), and is pursuing a PhD in systems engineering at the Naval Postgraduate School. He is a South Carolina Registered Professional Engineer and an INCOSE CSEP, and holds the OMG SysML Model Builder Fundamental Certification. [ryan.longshore@nps.edu]

Raymond Madachy, PhD—is a Professor in the Systems Engineering Department at the Naval Postgraduate School. His research interests include systems engineering tool environments for digital engineering, modeling and simulation of systems and software engineering processes, generative AI, and system cost modeling. He has developed cost estimation tools for systems and software engineering, and created the Systems Engineering Library (se-lib). His books include Software Process Dynamics, What Every Engineer Should Know about Modeling and Simulation, What Every Engineer Should Know about Python, and he co-authored Software Cost Estimation with COCOMO II and Software Cost Estimation Metrics Manual for Defense Systems. [rjmadach@nps.edu]

Abstract

Expert consensus is a critical component of decision-making in systems engineering, where stakeholder input and complex trade-offs must be carefully weighed. Traditionally, consensus-building techniques such as the Delphi Method, Nominal Group Technique (NGT), and Multi-Voting have been used to aggregate expert human opinions systematically. Constant lingering challenges prove to be deterrents such as time-intensive and extensive coordination efforts required to gather Subject Matter Experts (SMEs). With the advent of Large Language Models (LLMs), there exists the potential to capture the expert knowledge and leverage AI to streamline consensus-building.

This conceptual paper explores the feasibility of LLM-assisted consensus methods in the context of systems engineering. We evaluate consensus methods based on their structure, expert interaction requirements, and compatibility with LLMs, followed by identifying which methods could be enhanced through AI-driven automation. Through a comparative analysis, we hypothesize the methods best suited for LLM augmentation or full automation and explore their potential applications in systems engineering. Finally, we discuss future research directions for both AI-driven and hybrid human-AI consensus frameworks.

Keywords: Large Language Models (LLMs), Artificial Intelligence (AI), Consensus Methods, Systems Engineering, Feasibility Study

Introduction

The presence of accessible, capable AI systems has become widespread and presents a tool that should be leveraged intelligently as a force multiplier. The next generation of language models will require a shift from a "one size fits all" model to domain-specific models (Ling et al., 2024). The models can be trained on their own or fine-tuned from foundational

models. Foundational models are models trained on general knowledge. Conceptually, the research discussed herein focus on the interactions between and how to employ these domain-specific models. With each model trained on domain-specific knowledge, the similarity to being considered "AI Subject Matter Experts (SMEs)" as the same as having "human SMEs" starts to come to fruition.

The interaction between SMEs is a commonly orchestrated event for systems engineers. Systems engineers, acting as the glue between SMEs, sponsors, and project managers, are well poised to leverage domain-specific models in situations where a SME may not be available or too costly. This paper will explore the consensus methods commonly used by systems engineers for soliciting domain-specific knowledge to make informed decisions, discuss implementation architectures that are feasible for usage with language models, and propose systems engineering use cases and examine their challenges to implementation.

Overview of Consensus Methods

While language models have an inherent ability to synthesize large corpuses of information, their ability to come to a consensus among several models has been less studied, although interesting effects have been found at scale (Marzo et al., 2025). Some research has been done on hybrid consensus methods, including both humans and AI to come to a consensus (Chen et al., 2023; Fogliato et al., 2022; Hirosawa et al., 2024; Papakonstantinou et al., 2025; Punzi et al., 2024). Research has also been done on some of the challenges associated with hybrid consensus methods (Vaccaro et al., 2024).

The consensus methods considered are among some of the most common, including: the Delphi Method, the Fuzzy Delphi Method, Structured Expert Judgment (SEJ) also called Cooke's Method, Nominal Group Technique (NGT), the Stepladder Technique, Dialectical Inquiry, and Multi-Voting (Dot Voting). A summary table of each of these consensus methods' strengths and weaknesses is in Table 1, Consensus Methods Strengths and Weaknesses. A deeper dive into each consensus method follows. The sequence diagrams generated are intended to be representative of the most implementations of each technique, although there were slight variations present between different pieces of literature.

Table 1. Consensus Methods Strengths and Weaknesses

Method	Strengths	Weaknesses
Delphi Method	Reduces bias, allows for geographic dispersion, and provides a systematic approach to achieving consensus.	Time-consuming, lacks interaction, and may not achieve consensus.
Fuzzy Delphi Method	Captures ambiguity and uncertainty in expert opinions.	Complex for non-experts and requires fuzzy logic expertise.
Structured Expert Judgment	Provides quantitative outputs and handles uncertainty effectively.	Resource-intensive and requires expertise.
Nominal Group Technique	Encourages equal participation and produces clear prioritization.	Limited to small groups and time-consuming.
Multi-Voting	Quick, easy to implement, and provides clear prioritization.	May not capture nuances and can be influenced by voting strategies.
Stepladder Technique	Reduces dominance and improves decision quality.	Time-consuming and requires planning.
Dialectical Inquiry	Encourages critical thinking and creative solutions.	Contentious and may not achieve consensus.

The Delphi method is a structured, iterative process used to gather and consolidate expert opinions. It involves multiple rounds of questionnaires, with feedback provided to participants between rounds to encourage convergence of opinions. Experts respond to questionnaires in multiple rounds, with anonymous feedback usually in the form of the group average provided after each round. Consensus is typically defined as a percentage of agreement (e.g., 70–80%) or convergence variance of responses (e.g., +/- 1 on a ranking

scale). While this method attempts to reduce bias from dominant personalities through anonymous responses and the use of a facilitator, the process can be time-consuming and may not always achieve the set consensus threshold.

The authors have employed the Delphi method for estimating systems engineering cost model parameters for using AI (Madachy et al., 2025). Previously the Constructive Systems Engineering Cost Model (COSYSMO) and Constructive Cost Model (COCOMO) for software development were developed and calibrated with both expert judgment data via Delphi surveys and historical project data (Boehm et al., 2000; Valerdi, 2005). The Delphi method is also commonly used in clinical settings, among other domain-specific fields (Chan, 2022; Erffmeyer, 1981; Hutchings et al., 2006; Kauppi et al., 2023; Papakonstantinou et al., 2025; Spranger & Niederberger, 2025; Vedantham et al., 2023). The process is captured in Figure 1, Delphi Method Sequence Diagram.

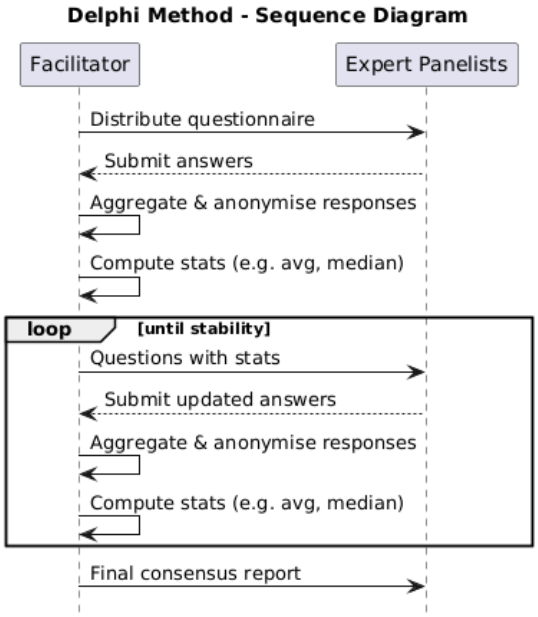

Figure 1. Delphi Method Sequence Diagram

The Fuzzy Delphi method integrates fuzzy logic with the traditional Delphi Method to capture the uncertainty in expert judgments. It is a structured, iterative process used to gather and consolidate expert opinions. It involves multiple rounds of questionnaires, with feedback provided to participants between rounds to encourage convergence of opinions. Experts still respond to questionnaires in multiple rounds, with anonymous feedback provided after each round, but use ranges—a fuzzy score—to compute convergence. This method lends itself best to situations where precise data is unavailable, but the learning curve is steep for facilitators new to fuzzy logic. The Fuzzy Delphi method is commonly used in situations where there is substantial ambiguity that needs to be quantified (Mohamad et al., 2015; Nayebpour & Sehhat, 2023; Padzil et al., 2021; Rahman & Kamauzaman, 2022; Rani et al., 2023). The process is captured in Figure 2, Fuzzy Delphi Method Sequence Diagram.

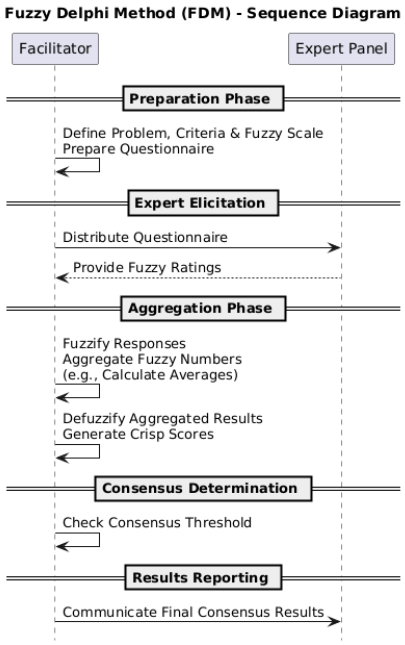

Figure 2. Fuzzy Delphi Method Sequence Diagram

The Structured Expert Judge (SEJ) method, also known as Cooke's method, uses expert opinions to quantify and produce probabilistic estimates. Each expert provides their individual assessment, all responses are aggregated, statistical weighting models are applied, and calibration is included if necessary. The quantitative output of the process is desirable, particularly for ambiguous and complex issues, although the process requires an expert to design. Cooke's method is commonly used within the nuclear field, ecosystems, and public health, among others (Colson & Cooke, 2018; Cooke et al., 2021; Felfernig & Le, 2023; Ullrika Sahlin, 2023). The process is captured in Figure 3, Structured Expert Judgment Sequence Diagram.

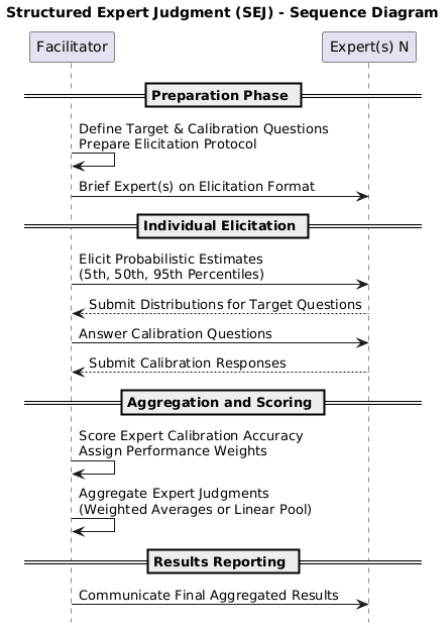

Figure 3. Structured Expert Judgment Sequence Diagram

The Nominal Group Technique (NGT) method is structured as a face-to-face consensus method that combines individual brainstorming with group discussion. NGT is primarily for structured idea generation and prioritization. Individuals brainstorm ideas, then ideas are then shared by each individual, one at a time while ideas are publicly recorded. Once all individuals have shared their ideas, the floor is open to group discussion with the focus on clarification of the ideas. Once clarifications are complete, everyone ranks the ideas that are most important or relevant. With humans, the process is typically limited to small groups, requires a facilitator, and can be time consuming. The output of this process is a list of ranked ideas, which can be used as inputs to other consensus methods to narrow down the list, such as the multi-voting method. Common usages include clinical studies and teaching, among others (Erffmeyer, 1981; Hutchings et al., 2006; Mousa et al., 2022; Rahman & Kamauzaman, 2022). The process is captured in Figure 4, Nominal Group Technique Sequence Diagram.

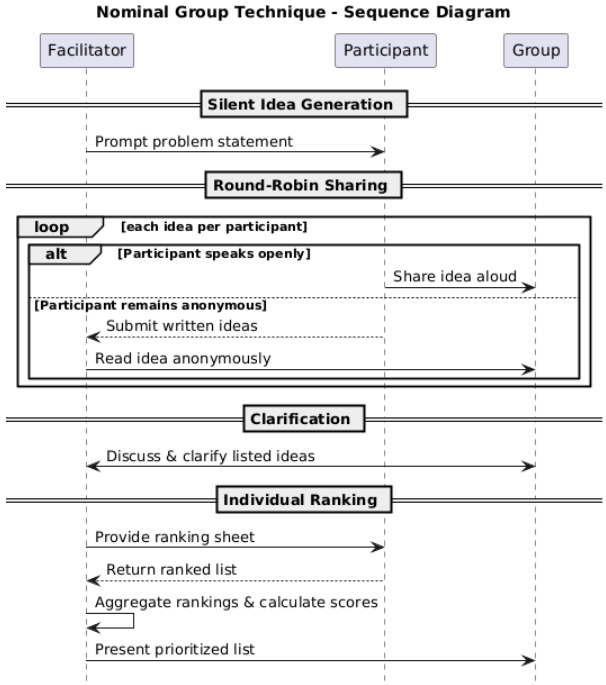

Figure 4. Nominal Group Technique Sequence Diagram

The Stepladder Technique is structured such that individual opinions are gradually added to the group discussion. In a tiered fashion, group sizes gradually increase. The process would start with individuals paired up who discuss their thoughts, followed by merging pairs to form small groups. Discussions continue. Small groups are then merged into a larger group. The process continues until all participants are in a single group. This method typically encourages participation from all members and reduces group think but requires a significant amount of time and is not typically suited for large groups (Rogelberg et al., 1992; Rogelberg & O'Connor, 1998).

The sequence diagram in Figure 5, Stepladder Sequence Diagram, presents a maximum of eight participants, although it could have as many as the facilitator or consensus designer would like and is merely a medium to communicate the process.

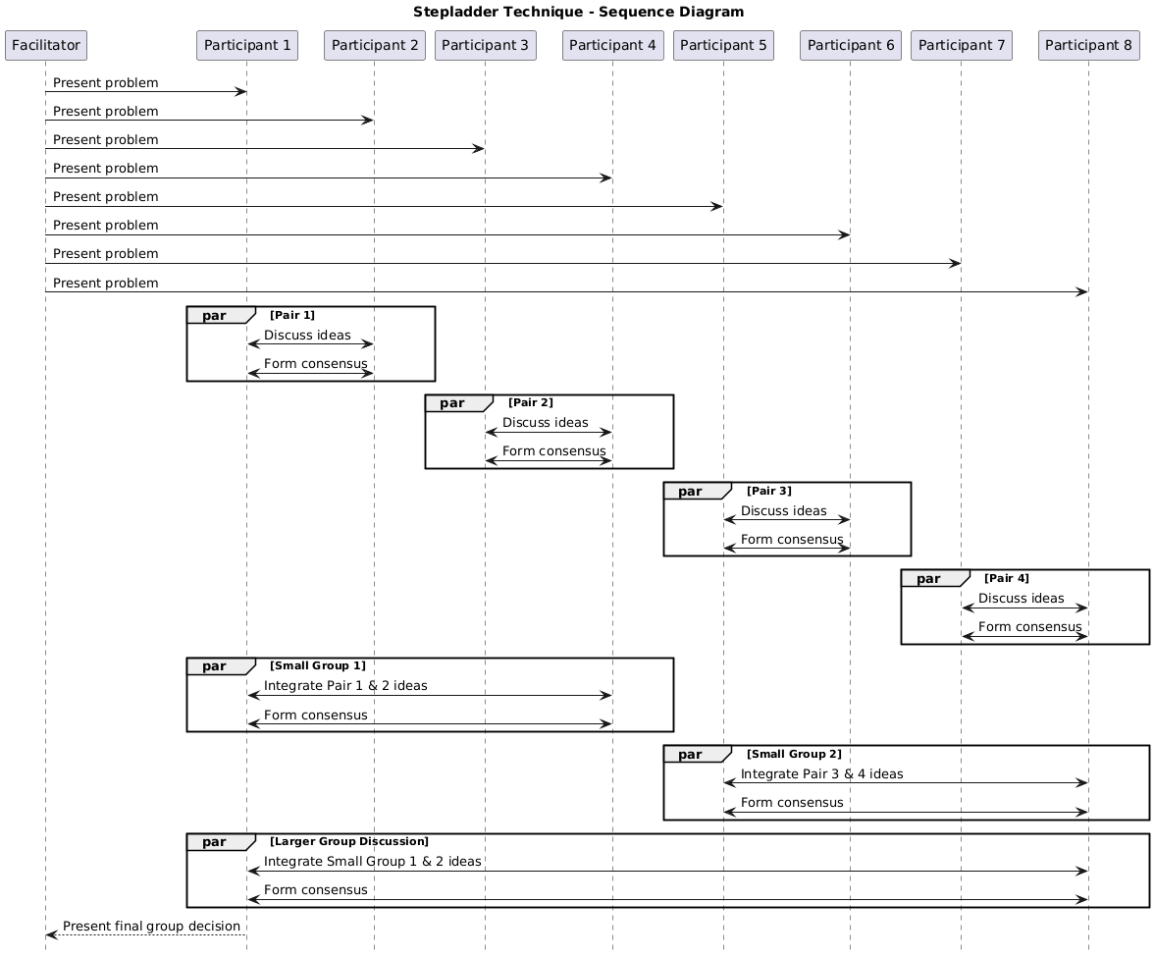

Figure 5. Stepladder Sequence Diagram

The Dialectical Inquiry method is focused on the premise of presenting opposing viewpoints to stimulate critical thinking among a group. Participants are divided up into groups that are assigned to argue for or against a proposition, and the debate continues until a consensus is reached. While the method can encourage creative solutions and foster new viewpoints, it can be contentious, requiring tactful facilitation, or, just as likely to ultimately not reach a consensus (Fjermestad, 1994; Priem & Price, 1991; Tung & Quaddus, 2001). The process is captured in Figure 6, Dialectical Inquiry Sequence Diagram.

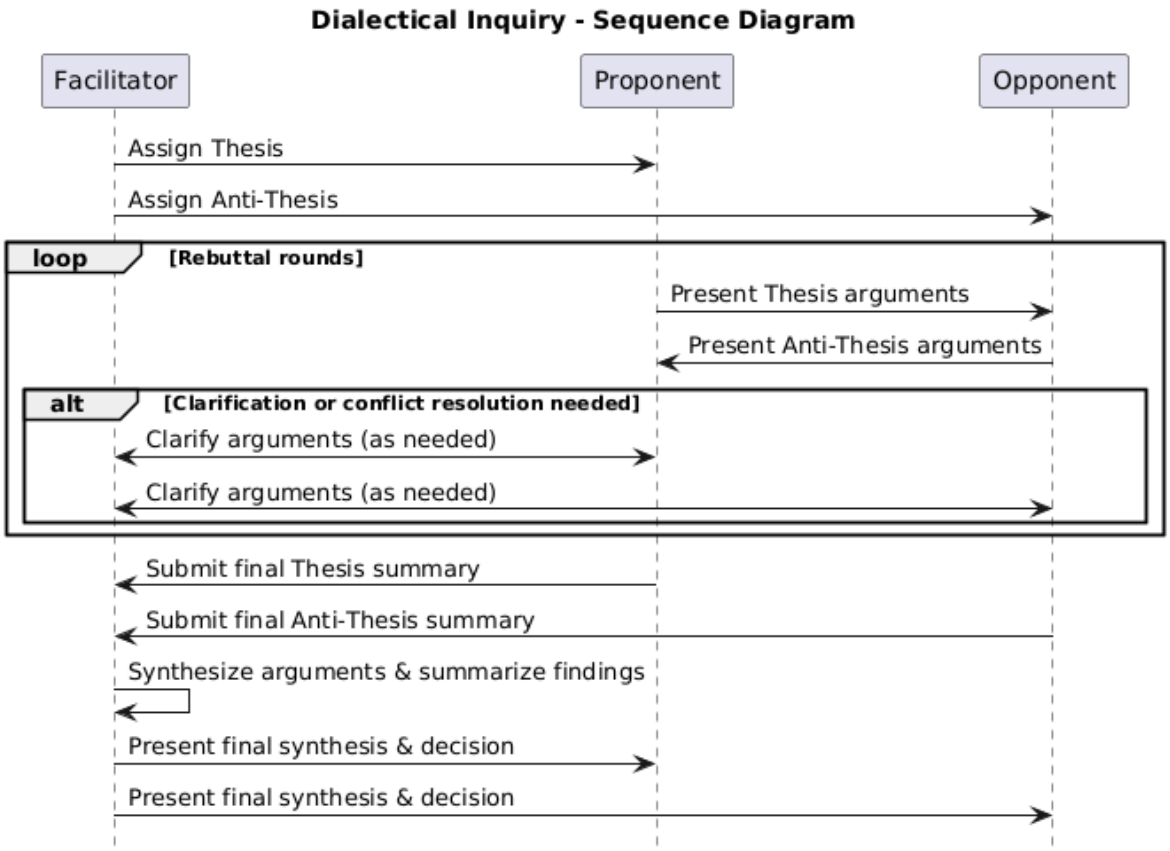

Figure 6. Dialectical Inquiry Sequence Diagram

The Multi-Voting method, also known as Dot Voting, is commonly used in the six-sigma process, where individuals vote on multiple items from a list. The list is generated ahead of time and may be generated via another consensus method like the NGT method. Multi-voting is then used to narrow down the list of options based on the group's consensus. Participants allocate their votes by either assigning a limited number (usually half the number of items) or by ranking all items on the list. The votes are then compiled and top items are presented. Multi-voting weighs every individual's vote equally. The selection process can become time-consuming or cumbersome for large lists. Multi-voting is used to narrow down a list of options as it is a simple voting mechanism used in Six Sigma practices and a variety of fields (American Society for Quality, 2025; Atlassian Community, 2024; Digital Healthcare Research, 2025; Hessing, 2015; Nielsen Norman Group, 2025). The process is captured in Figure 7, Multi-Voting Sequence Diagram.

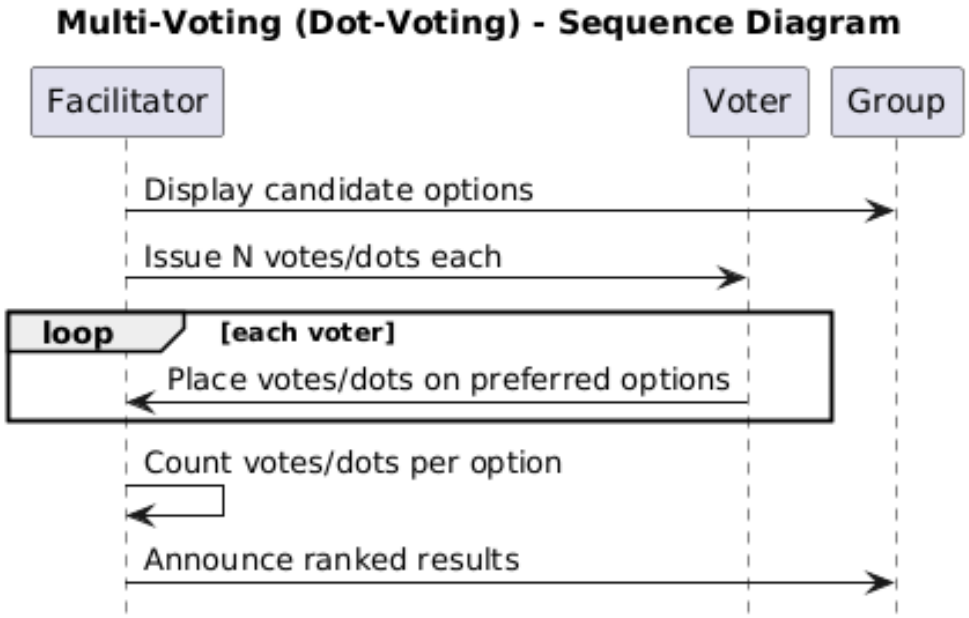

Figure 7. Multi-Voting Sequence Diagram

Summary of Consensus Method Characteristics

For each consensus method, a few characteristics were captured to support decision making for consensus method selection. The characteristics chosen are the columns found in Table 2, Consensus Methods Overview, and are anonymity, iteration, facilitation, output type, group interaction type, and the aggregation method.
Anonymity indicates whether participants provide input anonymously, which can affect group dynamics and bias mitigation. Possible values for this field include Yes, No, Partial, or Optional.

- Yes means full anonymity is maintained between participants.
- No means contributions are made openly.
- Partial means some anonymity exists during one or more of the stages of the consensus process.
- Optional means anonymity may or may not be used depending on the implementation style.

Iteration indicates whether the method includes repeated rounds of input and feedback, which can help refine judgments and converge on consensus. Possible values for this field include Single Round, Multiple Rounds, Built-in, or Optional.

- Single Round means the method is conducted in a single structured session without repetition.
- Multiple Rounds means the method explicitly involves repeated cycles of input, feedback, and revision.
- Built-in means iterative progression is inherently embedded in the method's structure.
- Optional means iteration is not required but can be included at the facilitator's discretion or based on group needs.

Facilitation refers to the level of structured guidance needed to execute the method effectively. Possible values for this field include Facilitator-Driven or Facilitator-Supported.

- Facilitator-Driven means a central facilitator is required to guide the process, manage feedback rounds, and enforce structure.
- Facilitator-Supported means a facilitator helps organize and maintain flow but does not drive every part of the process.

Output Type describes the nature of the results produced by the method. This determines whether the outcomes are narrative, numerical, or both, which influences how results are interpreted and used in decision-making. Possible values for this field include Qualitative, Quantitative, or Both.

- Qualitative means the outputs are primarily textual in nature.
- Quantitative means the outputs are numerical, such as rankings, vote tallies, or probabilistic scores.
- Both means the method can produce either qualitative insights or quantitative metrics.

Group Interaction Type identifies how participants communicate and collaborate during the method, which affects scheduling, group dynamics, and tool selection. Possible values for this field include Asynchronous or Synchronous.

- Asynchronous means participants provide input independently and at different times.
- Synchronous means participants interact in real time.

Decision Aggregation Method defines how individual participant inputs are synthesized into a collective judgment. This mechanism is central to reaching consensus or selecting preferred alternatives. Possible values for this field include None, Optional, Ranking, Count-Based, Scoring, or Fuzzy Scoring.

- None means there is no formal aggregation; consensus may emerge through discussion or argumentation.
- Optional means aggregation may or may not be used depending on context or facilitation style. The aggregation method could be one of the other methods but is not required.
- Ranking means participants order alternatives by preference, typically in descending importance.
- Count-based means options are selected or voted on with multiple tallies.
- Scoring means participants assign numeric ratings to options, which are then averaged or aggregated.
- Fuzzy Scoring means participants express uncertainty through ranges or fuzzy values (e.g., minimum, most likely, and maximum), which are aggregated using fuzzy logic methods.

Table 2. Consensus Methods Overview

Method	Anonymity	Iteration	Facilitation	Output Type	Group Interaction Type	Aggregation Method
Delphi	Yes	Multiple Rounds	Facilitator-Driven	Both	Asynchronous	Scoring
Fuzzy Delphi	Yes	Multiple Rounds	Facilitator-Driven	Quantitative	Asynchronous	Fuzzy Scoring
Structured Expert Judgment	Yes	Optional	Facilitator-Driven	Quantitative	Asynchronous	Scoring
Nominal Group Technique	Partial	Single Round	Facilitator-Driven	Both	Synchronous	Ranking
Multi-Voting	Optional	Optional	Facilitator-Supported	Quantitative	Synchronous	Count-based
Stepladder Technique	No	Built-in	Facilitator-Supported	Both	Both	Optional
Dialectical Inquiry	No	Multiple Rounds	Facilitator-Supported	Qualitative	Synchronous	None

The Implementation of Consensus Methods with Language Models

In order to understand the trade-offs between implementing consensus methods with language models, this research proposes criteria to qualitatively assess between the methods. A list of criteria was brainstormed to include: parallelizability, number of personas, agent persona archetypes, inter-agent communication pattern, and memory length.

The parallelizability criterion is how much of the method can be parallelized (e.g., agents working independently at the same time) where: High is fully parallel, Medium is some steps parallel, some sequential, and Low is mostly sequential. The number of AI personas is the recommended minimum number of distinct AI agents needed to implement the method. Agent persona archetypes are the types of roles or behavioral archetypes needed among the AI agents. The inter-agent communication pattern is how the AI agents exchange information during the process. Last but not least, memory length refers to how much dialogue or context history each agent needs to maintain during the method's execution where a single chat only requires one-off responses and conversational requires ongoing memory of turns or rounds.

Table 3, Consensus Method Implementation Characteristics, summarizes all of the evaluated criteria for each consensus method. In circumstances where a synthesizer persona is recommended, the role can be typically merged with the facilitator role, which is synonymous with the sequence diagrams. Some assumptions were made, including: 1. This is the logical formation of personas, but may be implemented as separate LLM calls or a single LLM stepping through roles and 2. If multi-round option is selected, this would be conversational.

Table 1 Consensus Method Implementation Characteristics

Consensus Method	Parallelizability	# AI Personas	Agent Persona Archetypes	Inter-Agent Communication Pattern[1]	Memory Length
Delphi Method	Medium	3+	Facilitator, Expert, Synthesizer	Hub-and-Spoke	Conversational
Fuzzy Delphi Method	Medium	3+	Facilitator, Expert, Synthesizer	Hub-and-Spoke	Conversational
Structured Expert Judgment	High	3+	Facilitator, Expert, Synthesizer	Hub-and-Spoke	Single Chat[2]
Nominal Group Technique	Low	4+	Facilitator, Creative Expert, Reasoning Expert, Summarizer	Group Broadcast	Conversational
Multi-Voting	High	2+	Faciliator, Participant	Blind Broadcast	Single Chat
Stepladder Technique	Medium	3+	Faciliator, Participant, Synthesizer	Progressive Entry	Conversational
Dialectical Inquiry	Low	3+	Facilitator, Thesis Supporter, Antitheses Supporter	Sequential Debate	Conversational

The agent persona archetypes are major roles, including the facilitator, expert, synthesizer, creative expert, reasoning expert, participant, thesis supporter, and antithesis supporter. Some of these personas could be merged under certain circumstances, like the facilitator and synthesizer. In general, the following purpose of each of these archetypes is

- Facilitator: Guides the process, enforces rules, moderates the flow. Also known as the Conductor within agentic frameworks.

- Expert: Provides substantive technical input or judgment. A generalization of a creative or reasoning expert.
- Synthesizer: A decomposition of the facilitator role to summarize and aggregate information.
- Creative Expert: An expert that focuses on brainstorming new ideas in early stages.
- Reasoning Expert: An expert that focuses on substantiating, prioritizing, or ranking options.
- Participant: A general contributor without major specialization.
- Thesis Supporter: Defends an assigned position.
- Antithesis Supporter: Critiques the thesis with counter-arguments.

Each of these persona archetypes generally values a different level on the "creativity" scale, which is synonymous with temperature for language models. The relationship between temperature and persona archetype is continued in Table 4, Agent Persona Temperatures.

Table 4. Agent Persona Temperatures

Persona Archetype	Suggested Temperature	Rationale
Facilitator	Low	Must keep structure, restate prompts, and remain neutral. Deterministic output avoids accidental bias or drift.
Expert	Low–Medium	Needs factual depth with little room for nuance or hypothesis generation. Too much randomness risks misinformation; too little may freeze creative problem solving.
Synthesizer / Summarizer	Low	Primary duty is faithful condensation. Higher temperature could invent facts or reorder logic.
Creative Expert	Medium–High	Charged with idea generation. Higher temperature encourages novel alternatives and divergent thinking.
Reasoning Expert	Low–Medium	Focus on logical evaluation; moderate temperature keeps reasoning flexible but still disciplined.
Participant	Low–Medium	Casting or explaining a preference benefits from mild variability (tie break rationales) but must stay consistent with criteria.
Thesis Supporter / Antithesis Supporter	Medium–High	Goal is vigorous argumentation. Higher temperature produces persuasive rhetoric, counter examples, and creative rebuttals, effectively fueling dialectical tension.

The characteristics identified and qualities assessed will help with adaptation into agentic frameworks like CrewAI, Autogen, OpenAI's swarm, among many others from the open source community (GitHub, 2024; Microsoft, 2023; n8n.io, 2025; OpenAI, 2024; SuperAGI, 2025).

Systems Engineering Applications of LLM-Centric Consensus Methods

Consensus plays a critical role in systems engineering by ensuring that the boundaries of complex technical trade spaces reflect the collective judgment of multidisciplinary stakeholders such that the systems engineer can make an informed decision. In some cases, expert consensus enables the reconciliation of conflicting priorities—such as cost, performance,

schedule, and safety—through structured deliberation while in other cases, consensus simply populates the bounds of the trade space. As systems grow in complexity with exponential interdependencies, the ability to achieve consensus among domain-specific SMEs becomes a cornerstone of successful systems engineering practices.

Language models are emerging as powerful AI tools for decision support in systems engineering. Given their capability for synthesizing large swaths of information and offering structured insights, they are a natural support tool for systems engineers. When integrated into tools like Model-Based Systems Engineering (MBSE) environments with SysMLv2 textual notation, LLMs can be just as aware as the systems engineer, with hopes of enhancing traceability by cross-referencing system artifacts. In the Department of Defense (DoD), domain-specific knowledge bases are plentiful. Connecting a language model to these domain-specific knowledge bases and simply having the models interact rather than the full exchange of data is desirable for compartmentalization reasoning and security. The ability to have these domain-specific aware, black box models and having an interplay between them may bring a level of consensus on complex topics not before made. To further elaborate on the application of consensus methods in systems engineering, two use cases were chosen to pontificate on how these LLM-centric methods would apply to common systems engineering problems.

The first use case presented is one pertaining to requirements engineering, specifically stakeholder requirements solicitation—arguably the most important stage—where we can use several AI agents with varying personas to brainstorm pertinent stakeholder needs to requirements, followed by a consensus method for pruning this large list into a pruned prioritized list of stakeholder requirements.

We hypothesize that using NGT or the stepladder technique for an initial pass at requirements are both good approaches. In this specific case, we propose that NGT is used for its ability to brainstorm from many viewpoints, followed by an optional multi-voting method for pruning the list should the list be too long. The final pruning would need to be guided by a human, but the facilitation can still occur from an AI persona. This scenario is about surfacing what matters to diverse users. The goal is to ensure each voice is heard and that the initial capability list is representative, even if imperfect or perhaps lacking technical rigor, depending on the personas selected. Personas might include all typical SMEs from an Integrated Product Team (IPT), such as but not limited to mechanical, electrical, structural, aerospace, logistics, and program management.

The second use case presented is about performing risk analysis. Generally, risk analysis requires somewhat specialized knowledge like fault tree analyses, FMEA, or FMECA. In this scenario, it is assumed that the brainstorming phase has been conducted and the focus is on narrowing down the most plausible solution(s).

We hypothesize that given the typically required specialized knowledge, consensus methods that leverage experts like SEJ or a Delphi approach are appropriate. Both methods enable the voice of the experts to be heard with optional iterative feedback between experts. Using the knowledge available to the agents, the question posed would be for a ranking of severity. Personas might include a level of sub-field specificity like reliability engineers or availability engineers instead of more general personas like mechanical, electrical, or similar.

Challenges to Implementation

There are two main categories of challenges to implementation: challenges that are inherent to LLMs in general and challenges that are inherent to the consensus framework used. Challenges like accuracy, bias, role consistency, human diversity at the single model level, and memory are all LLM-inherent challenges with AI technology. When it comes to implementing the

consensus framework, things like emulating human diversity where at the group level emergent behavior could be present, iteration management, human oversight of the process, and consensus metrics are all present. The challenges are summarized in Table 5, Challenges to Implementation.

Table 5. Challenges to Implementation

Challenge Category	Challenge	Considerations
LLM-Inherent	Accuracy and bias	RAG, Fine-tuning, benchmarked models
LLM-Inherent	Role consistency	Role templates, temperature tuning
Both	Emulating human diversity	Model/temperature mixing, diverse personas, prompt engineering
Both	Memory and iteration management	Long-context models, vector stores, iterative summarization
Consensus Framework	Human oversight	Human-on-the-loop, checkpoints, audit logs
Consensus Framework	Consensus and convergence metrics	Entropy/rubric metrics, semantic convergence checks, human final judgment

Future Work and Research Opportunities

There is a plethora of future research available in this area—the most obvious is the implementation of each of the consensus methods discussed herein into open source tooling, such as a Python library to be used with language models. The library could include human in the loop versus human on the loop interfaces as well as support for hybrid consensus framework structures for human-AI teaming.

The application of any of these consensus frameworks with AI to a typical systems engineering use case as discussed would inform practitioners about the usefulness of using AI as a force multiplier. Multi-modal or vision-models could be used to assess prototype photographs of systems. Convergence rates for varying temperature or “creativity” levels could be investigated. The ability to scale participants to levels incapable of human participation also warrants investigation, with the hopes of finding emergent behavior not previously possible at scale. For example, the application of large scale ranking without reaching cognitive overload of participants with 100s of items to prioritize.

Conclusion

The usage of consensus methods with AI necessitates further research. The systems engineering field would benefit greatly from gathering consensus from multiple language models across different phases of the systems engineering life cycle. Practicing systems engineers and SMEs could supplement their own knowledge bases with AI personas to enhance viewing problems from different perspectives.

As AI tool suites continue to propagate, systems engineers need to consider the assimilation of AI tools with classical methods of reaching decisions via consensus methods with experts. The ability to process, understand, and make informed decisions within a trade space is only going to become more challenging for systems engineers as systems of systems continue to become more complex. Domain-specific AI models can help relieve some of the complexity—from understanding an entire model-based systems engineering (MBSE) model in SysMLv2 textual form to understanding entire knowledge bases of domain-specific data, gathered from decades of expert practice in the field.

The future of systems engineering will not depend solely on more powerful AI models, but on how effectively humans and machines collaborate. The challenge lies in engineering decision frameworks that balance trust, skepticism, and synthesis across diverse AI and human perspectives for effective, practical implementation.

Acknowledgements

This work has benefited from the use of generative AI tools, including ChatGPT, Claude, and SciSpace, for brainstorming, writing assistance, and code development (Anthropic AI, 2025; OpenAI, 2025; SciSpace, 2025). These tools were employed to enhance conceptual clarity, improve code efficiency, and support literature synthesis. All AI-generated outputs were critically reviewed, refined, and validated to ensure accuracy and alignment with academic integrity. Their contributions were limited to supporting the research process, and final responsibility for the content, analysis, and conclusions remains with the authors.

References

American Society for Quality. (2025). *What is multivoting? NGT voting, nominal prioritization*. https://asq.org/quality-resources/multivoting?srsltid=AfmBOoqNNFsEpIlk7pjC74Reu9AhCE_GBNM1Fb8U1wPFElgWNi9yXlKC

Anthropic AI. (2025). *Anthropic home*. https://www.anthropic.com/

Atlassian Community. (2024, March 27). *Dot voting: An effective way for group decision-making*. https://community.atlassian.com/forums/App-Central-articles/Dot-Voting-An-Effective-Way-for-Group-Decision-Making/ba-p/2653536

Boehm, B., Abts, C., & Chulani, S. (2000). Software development cost estimation approaches – A survey. *Annals of Software Engineering*, *10*(1–4), 177–205. https://doi.org/10.1023/A:1018991717352

Chan, P. (2022). An empirical study on data validation methods of delphi and general consensus. *Data*, *7*(2), 18. https://doi.org/10.3390/data7020018

Chen, X., Xu, X., Pan, B., & Zhang, W. (2023). *Human-machine collaboration-driven consensus method for large-scale group decision-making based on reinforcement learning algorithm and its application*. https://doi.org/10.2139/ssrn.4623328

Colson, A. R., & Cooke, R. M. (2018). Expert elicitation: Using the classical model to validate experts' judgments. *Review of Environmental Economics and Policy*, *12*(1), 113–132. https://doi.org/10.1093/reep/rex022

Cooke, R. M., Marti, D., & Mazzuchi, T. (2021). Expert forecasting with and without uncertainty quantification and weighting: What do the data say? *International Journal of Forecasting*, *37*(1), 378–387. https://doi.org/10.1016/j.ijforecast.2020.06.007

Digital Healthcare Research. (2025). *Multivoting*. https://digital.ahrq.gov/health-it-tools-and-resources/evaluation-resources/workflow-assessment-health-it-toolkit/all-workflow-tools/multivoting

Erffmeyer, R. (1981). *Decision-making formats: A comparison on an evaluative task of interacting groups, consensus groups, the nominal group technique, and the Delphi technique.* [Doctor of Philosophy, Louisiana State University and Agricultural & Mechanical College]. https://doi.org/10.31390/gradschool_disstheses.3593

Felfernig, A., & Le, V. M. (2023). An overview of consensus models for group decision-making and group recommender systems. *User Modeling and User-Adapted Interaction*. https://doi.org/10.1007/s11257-023-09380-z

Fjermestad, J. L. (1994). *Group strategic decision-making in a computer-mediated-communications environment: A comparison of dialectical inquiry and constructive consensus approaches*.

Fogliato, R., Chappidi, S., Lungren, M. P., Fitzke, M., Parkinson, M., Wilson, D. U., Fisher, P., Horvitz, E., Inkpen, K., & Nushi, B. (2022). Who goes first? Influences of human-AI workflow on decision making in clinical imaging. *2022 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3531146.3533193

GitHub. (2024). *CrewAI*. https://github.com/crewAIInc/crewAI

Hessing, T. (2015, November 21). Multivoting. *Six Sigma Study Guide*. https://sixsigmastudyguide.com/multivoting/

Hirosawa, T., Shiraishi, T., Hayashi, A., Fujii, Y., Harada, T., & Shimizu, T. (2024). Adapting artificial intelligence concepts to enhance clinical decision-making: A hybrid intelligence framework. *International Journal of General Medicine*, *17*, 5417–5422. https://doi.org/10.2147/ijgm.s497753

Hutchings, A., Raine, R., Sanderson, C., & Black, N. (2006). A comparison of formal consensus methods used for developing clinical guidelines. *Journal of Health Services Research & Policy*, *11*(4), 218–224. https://doi.org/10.1258/135581906778476553

Kauppi, K., Roos, E., Borg, P., & Torkki, P. (2023). Building consensus on domains of wellness using Finnish and international expert panels: A Delphi-method study. *American Journal of Health Promotion*, 8901171231204147. https://doi.org/10.1177/08901171231204147

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Mehta, D., Pasquali, S., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., … Zhao, L. (2024). *Domain specialization as the key to make large language models disruptive: A comprehensive survey* (No. arXiv:2305.18703). arXiv. https://doi.org/10.48550/arXiv.2305.18703

Madachy, R., Bell, R., & Longshore, R. (2025). A generative AI-driven systems engineering maturity and cost modeling framework. *Proceedings of the 2025 Conference on Systems Engineering Research*.

Marzo, G. D., Castellano, C., & Garcia, D. (2025). *AI agents can coordinate beyond human scale* (No. arXiv:2409.02822). arXiv. https://doi.org/10.48550/arXiv.2409.02822

Microsoft. (2023). *AutoGen*. GitHub. https://github.com/microsoft/autogen

Mohamad, S. N. A., Embi, M., & Nordin, N. (2015). Determining e-Portfolio elements in learning process using fuzzy Delphi analysis. *International Education Studies*, *8*, 171–171. https://doi.org/10.5539/ies.v8n9p171

Mousa, M., Teede, H. J., Garth, B., Winship, I., Prado, L., & Boyle, J. (2022). Using a modified Delphi approach and nominal group technique for organisational priority setting of evidence-based interventions that advance women in healthcare leadership. *International Journal of Environmental Research and Public Health*, *19*(22), 15202. https://doi.org/10.3390/ijerph192215202

n8n.io. (2025). *n8n: Workflow automation tool*. GitHub. https://github.com/n8n-io/n8n

Nayebpour, H., & Sehhat, S. (2023). Designing the competency model of human resource managers based on paradox theory (Case study: Information and communication technology industry). *International Journal of Organizational Analysis*, *32*. https://doi.org/10.1108/IJOA-02-2023-3645

Nielsen Norman Group. (2025). *Dot voting: A simple decision-making and prioritizing technique in UX*. https://www.nngroup.com/articles/dot-voting/

OpenAI. (2024). *Swarm*. GitHub. https://github.com/openai/swarm

OpenAI. (2025). *OpenAI*. Ask ChatGPT Anything. https://openai.com/

Padzil, M. R., Karim, A., & Husnin, H. (2021). Employing DDR to design and develop a flipped classroom and project based learning module to applying design thinking in design and technology. *International Journal of Advanced Computer Science and Applications*, *12*. https://doi.org/10.14569/IJACSA.2021.0120988

Papakonstantinou, N., Van Bossuyt, D., Bell, R., Longshore, R., & Heikkila, M. (2025, January). *PrivateAIDELPHI: Adopting and adapting private AI for risk assessment of safety critical systems*. RAMS, Miramar Beach, FL. https://doi.org/10.1109/RAMS48127.2025.10935226

Priem, R. L., & Price, K. H. (1991). Process and outcome expectations for the dialectical inquiry, devil's advocacy, and consensus techniques of strategic decision making: *Group & Organization Management*, *16*(2), 206–225. https://doi.org/10.1177/105960119101600207

Punzi, C., Pellungrini, R., Setzu, M., Giannotti, F., & Pedreschi, D. (2024). AI, meet human: Learning paradigms for hybrid decision making systems. *arXiv.Org*, *abs/2402.06287*. https://doi.org/10.48550/arxiv.2402.06287

Rahman, N. H. N. A., & Kamauzaman, T. H. T. (2022). Developing key performance indicators for emergency department of teaching hospitals: A mixed fuzzy Delphi and nominal group technique approach. *Malaysian Journal of Medical Science*, *29*(2), 114–125. https://doi.org/10.21315/mjms2022.29.2.11

Rani, N. B. A., Hashim, M. E. A. B., Mustafa, W. A., Idris, M. Z. B., Al-Jawahry, H. M., & Ramadan, G. M. (2023). *Applying fuzzy Delphi method (FDM) to obtain the expert consensus in aesthetic experience (AX) and immersive experience (IX) elements for virtual reality historical event (VR historical event)*. 1–4. https://doi.org/10.1109/icmnwc60182.2023.10435812

Rogelberg, S. G., Barnes-Farrell, J. L., & Lowe, C. A. (1992). The stepladder technique: An alternative group structure facilitating effective group decision making. *Journal of Applied Psychology*, *77*(5), 730–737. https://doi.org/10.1037/0021-9010.77.5.730

Rogelberg, S. G., & O'Connor, M. S. (1998). Extending the stepladder technique: An examination of self-paced stepladder groups. *Group Dynamics: Theory, Research, and Practice*, *2*(2), 82–91. https://doi.org/10.1037//1089-2699.2.2.82

SciSpace. (2025). *SciSpace*. https://typeset.io/

Spranger, J., & Niederberger, M. (2025). How Delphi studies in the health sciences find consensus: A scoping review. *Systematic Reviews*, *14*(1). https://doi.org/10.1186/s13643-024-02738-3

SuperAGI. (2025). *SuperAGI: Autonomous AI agent framework*. GitHub. https://github.com/TransformerOptimus/SuperAGI

Tung, L. L., & Quaddus, M. (2001). Conflict management in dialectical inquiry, devil's advocacy and consensus-based decision making approaches in a GSS environment. *Proceedings of the Pacific Asia Conference on Information Systems (PACIS)*. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1094&context=pacis2001

Ullrika Sahlin (Director). (2023, February 7). *Structured expert judgment to assess uncertainty* [Video recording]. Lund University. https://www.youtube.com/watch?v=QPrawReZBxI

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Valerdi, R. (2005). *The constructive systems engineering cost model (COSYSMO)*.

Vedantham, S., Gloviczki, P., Carman, T. L., Schneider, O., Sabri, S. S., & Kolluri, R. (2023). Delphi consensus on reporting standards in clinical studies for endovascular treatment of acute iliofemoral venous thrombosis and chronic iliofemoral venous obstruction. *Circulation-Cardiovascular Interventions*, *16*, e012894. https://doi.org/10.1161/CIRCINTERVENTIONS.123.012894

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net