1. **Inter-Annotator Agreement Analysis: Interpretation**
2. **Overview of Agreement Levels**

The Cohen's Kappa analysis across all annotation dimensions reveals consistently **slight agreement** (κ = 0.001-0.082) among the four expert annotators, according to the Landis & Koch (1977) interpretation scale. This finding is significant for understanding the inherent challenges in subjectively evaluating image accessibility for people with cognitive disabilities.

1. **Dimension-Specific Analysis**

**Text-Image Alignment** emerged as the dimension with the highest inter-annotator agreement (κ = 0.073-0.082), suggesting that experts have relatively more consistent standards when evaluating whether images appropriately support and represent the accompanying simplified text. This finding is encouraging as text-image alignment is crucial for accessibility, indicating that with proper guidelines, experts can achieve reasonable consensus on this critical aspect.

**Text-related dimensions** (Text Simplicity: κ = 0.055-0.076; Text Quality: κ = 0.061-0.071) showed moderate consistency within the "slight" category. The similar agreement levels between text simplicity and quality suggest that experts have somewhat aligned perspectives on textual accessibility, likely due to more established criteria for evaluating plain language and readability.

**Image Quality** demonstrated the lowest agreement (κ = 0.001-0.021), approaching near-random consensus. This extremely low agreement highlights the highly subjective nature of aesthetic and technical quality judgments, suggesting that "quality" perceptions vary dramatically among experts and may require more objective, measurable criteria.

**Image Simplicity** (κ = 0.040-0.074) and **Ethics** (κ = 0.055-0.068) showed intermediate agreement levels within the slight category. The moderate consistency in image simplicity evaluation suggests that while experts can somewhat agree on visual complexity, significant subjectivity remains in interpreting what constitutes "simple" or "clear" visual content for cognitive accessibility.

1. **Sample Size Effect Analysis**

Interestingly, agreement levels remained remarkably stable across different sample sizes:

* **19 images** (all 4 experts): κ = 0.061 ± 0.027
* **30 images** (3+ experts): κ = 0.054 ± 0.022
* **113 images** (2+ experts): κ = 0.054 ± 0.019

This consistency suggests that the observed agreement patterns are robust and not artifacts of small sample sizes. The decreasing standard deviation with larger samples indicates more stable estimates, supporting the reliability of our findings.

1. **Expert Pair Variations**

Notable patterns emerged in pairwise comparisons:

* **Expert K and Expert L** consistently showed the highest agreement across most dimensions, particularly in text-related evaluations (κ up to 0.233 for text simplicity)
* **Expert A and Expert M** frequently exhibited lower agreement, suggesting potentially different annotation philosophies or interpretation of guidelines
* Some expert pairs showed negative kappa values (particularly in image quality), indicating agreement worse than chance, which highlights fundamental disagreements in evaluation criteria

1. **Implications for Accessibility Research**

**Methodological Implications**: The slight agreement levels, while low by some standards, are not uncommon in subjective annotation tasks involving aesthetic, cognitive, or accessibility judgments. These results underscore the inherent challenge of creating objective standards for evaluating content designed for people with cognitive disabilities, where individual differences in cognitive processing can vary significantly.

**Practical Implications**: The findings suggest that current evaluation frameworks for image accessibility may benefit from:

1. **More specific operational definitions** for each dimension
2. **Calibration training** for annotators to align evaluation standards
3. **Hybrid approaches** combining expert judgment with automated metrics
4. **Focus on consensus-building** around dimensions with higher agreement potential

**Research Validity**: Despite the slight agreement, the consistency across sample sizes and the meaningful patterns in dimension-specific variations support the validity of the annotation framework. The results provide valuable insights into which aspects of image accessibility can be more reliably evaluated and which require further methodological development.

1. **Contextualizing the Results**

These kappa values align with similar studies in accessibility and subjective content evaluation, where achieving substantial agreement (κ > 0.60) is challenging due to the inherently subjective nature of the task. The slight agreement observed here establishes a baseline for future research and highlights the need for continued refinement of accessibility evaluation methodologies.

The analysis demonstrates that while perfect agreement among experts is unrealistic for complex accessibility judgments, systematic patterns in disagreement can inform the development of more robust evaluation frameworks and training protocols for future accessibility assessments.