# CONDITION 1: NO ANCHOR PROMPTS

**DO NOT RESPOND WITH 'I'm sorry...' was added to features where the model consistently did not give an answer.**

**Model was given the prompt text, then “This is the rock you will label.” followed by the rock image.**

"This is the rock you will label."})

images\_content.append({"type": "image\_url", "image\_url": f"data:image/jpeg;base64,{base64\_image}"})

images\_content = []

*# Append the main image*

images\_content.append({"type": "text", "text": f"This is the rock you will label."})

images\_content.append({"type": "image\_url", "image\_url": f"data:image/jpeg;base64,{base64\_image}"})

payload = {

"model": "gpt-4-vision-preview",

"messages": [

{

"role": "user",

"content": [

{"type": "text", "text": prompt\_text}

] + images\_content

},

],

"max\_tokens": 300,

"temperature": 0.0

}

## Lightness

"In this trial, you will rate a rock on its darkness/lightness of color. A dark rock should receive a rating of 1.00 or 2.00. A light rock should receive a rating of 8.00 or 9.00. A rock that is medium in darkness/lightness should receive a medium rating. In some cases, parts of the rock may be light and other parts may be dark. In those cases, do your best to rate the 'average' lightness of the entire rock. Please try to use the full scale from 1.00 (darkest) through 9.00 (lightest) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.** "

## Grain Size

"In this trial, you will rate a rock on its average grain size. Rocks with no visible grain should receive a rating of 1.00 or 2.00. Rocks with an extremely coarse and fragmented grain should receive a rating of 8.00 or 9.00. Rocks with a medium grain should receive medium ratings. In some cases, parts of the rock may have a fine grain and other parts may have a coarse grain. In those cases, do your best to rate the 'average' grain size of the entire rock. Please try to use the full scale from 1.00 (no visible grain) through 9.00 (very coarse grain) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range. DO NOT RESPOND WITH 'I'm sorry...'**"

## Roughness

'In this trial, you will rate a rock on how smooth versus rough it appears to be. Rocks that appear to be very smooth should receive a rating of 1.00 or 2.00. Rocks that appear to be very rough should receive a rating of 8.00 or 9.00. Rocks that are medium in their smoothness/roughness should receive medium ratings. In some cases, parts of a rock may be smooth and other parts may be rough. In those cases, do your best to rate the "average" roughness of the entire rock. Please try to use the full scale from 1.00 (smoothest) through 9.00 (roughest) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.** '

## Shininess

'An object is "shiny" if it reflects light and is glossy. Note that dark-colored objects can still be shiny. In this trial, you will rate a rock on how dull versus shiny it appears to be. Rocks that appear to be very dull should receive a rating of 1.00 or 2.00. Rocks that appear to be very shiny and glossy should receive a rating of 8.00 or 9.00. Rocks that are medium in their dullness/shininess should receive medium ratings. In some cases, parts of a rock may be dull and other parts may be shiny. In those cases, do your best to rate the "average" shininess of the entire rock. Please try to use the full scale from 1.00 (dullest) through 9.00 (shiniest) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.**'

## Organization

'Some rocks have components that are very regular and organized, such as systematic layers, bands, or grains. Other rocks seem very disorganized, such as those with fragments that are glued together in haphazard fashion. In this trial, you will rate a rock on how disorganized versus organized it appears to be. Rocks that are very disorganized should receive a rating of 1.00 or 2.00. Rocks that are very organized should receive a rating of 8.00 or 9.00. Rocks that are medium in their organization, or that have no visible texture to rate, should receive medium ratings. Please try to use the full scale from 1.00 (most disorganized) through 9.00 (most organized) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.** '

## Chromaticity

'In this trial, you will rate a rock in terms of whether it has no color, cool color, or warm color. Rocks with no color (absolute black, gray or white) should receive a rating of 1.00 or 2.00. Rocks with cool colors (blue, blue/green, and green) should receive medium ratings (4.00, 5.00, or 6.00). Rocks with very warm colors (yellow, orange, red) should receive ratings of 8.00 or 9.00. Please try to use the full scale from 1.00 (no color) through 9.00 (warmest color) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.** '

## Red/Green Hue

"In this experiment, you will be presented with a picture of a rock. We would like you to rate the rock picture on a red-green contrast. Rocks that are most strongly red should receive ratings of 1.00 or 2.00. Rocks that are most strongly green should receive ratings of 8.00 or 9.00. Neutral rocks (black or white) that are absent of color should receive ratings of 5.00. For the remaining rocks, just decide whether the main color tends to be closer to red versus green. For example, most would agree that orange is closer to red than to green, so you might give orange rocks ratings of 2.00, 3.00, or 4.00. Likewise, most would agree that blue is closer to green than to red, so you might give blue rocks ratings of 6.00, 7.00, or 8.00. Please try to use the full scale from 1.00 to 9.00 in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range. DO NOT RESPOND WITH 'I'm sorry...'**"

## Porphyritic Texture

"In this experiment you will be presented with a picture of a rock. We are interested in your judgments about a very specific property of some of the rocks -- Certain kinds of rocks contain small fragments or pebbles that are glued into a separate background texture. THESE SMALL FRAGMENTS OR PEBBLES ARE SEPARATE FROM THE REST OF THE ROCK'S BACKGROUND ITSELF. We want you to rate each rock picture for this property. Rocks with no small fragments or pebbles glued into their separate background should receive a rating of 1.00 or 2.00. Rocks that definitely have small fragments or pebbles glued into their separate background should receive a rating of 8.00 or 9.00. Many rocks may be unclear cases; Some may have a coarse grain throughout, but don't really have separate small fragments glued into them. Other rocks may also be hard to judge because they have changes in shading that are not really separate glued fragments. These unclear cases should receive ratings of 4.00, 5.00 or 6.00. Please try to use the full scale from 1.00 through 9.00 in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range. DO NOT RESPOND WITH 'I'm sorry...'**"

## Pegmatitic Structure

"In this experiment you will be presented with a picture of a rock. Certain rocks have very large-sized crystals that are embedded in a SEPARATE background. The crystals will often (but not always) appear as large shiny bands. Your job in this experiment is simply to judge the extent to which the rock shown in each picture has this property. Rocks that have nothing like this property should receive ratings of 1.00 or 2.00. Rocks that have a hint of this property should receive ratings of 4.00, 5.00, or 6.00. Rocks that strongly display this property should receive ratings of 8.00 or 9.00. Please try to use the full scale from 1.00 through 9.00 in making your rating. Note: Because this property is very rare, most of the time your response will be between 1.00 and 2.00. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range. DO NOT RESPOND WITH 'I'm sorry...'**"

## Conchoidal Fracture

"In this experiment, you will be presented with a picture of a rock. We are interested in your judgments about a very specific property of some rocks. The property is called CONCHOIDAL FRACTURES. Conchoidal fractures are formed when pieces of a brittle rock chip off and leave behind smooth, curved surfaces resembling the inside of a seashell. Conchoidal fractures are typically found in glassy or fine-grained rocks. In this trial of the experiment, you will be shown a rock picture. We want you to rate the rock picture for the extent to which it has conchoidal fractures. Rocks with flat or jagged surfaces, or rocks with no fractures should receive a rating of 1.00 or 2.00. Rocks with smooth, curved indents or fractures resembling the inside of a seashell should receive a rating of 8.00 or 9.00. Many rocks may be unclear cases: Some rocks may have fractures where pieces of the rock were chipped off, but they may not be as smooth or curved as true conchoidal fractures. Other rocks may also be hard to judge because they have changes in shading or color. These unclear cases should receive ratings of 4.00, 5.00, or 6.00. Ratings of 8.00 or 9.00 should be given only for rocks for which you are absolutely sure they have conchoidal fractures. Most rocks do not have conchoidal fractures and should receive low ratings. Please try to use the full scale from 1.00 through 9.00 in making your ratings. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range. DO NOT RESPOND WITH 'I'm sorry...'**"

# CONDITION 2: ANCHOR PROMPTS

A line mentioning example images was added to the prompt e.g., An example of a very dark rock, a rock that is medium in darkness/lightness, and a very light rock is shown.

**Model was given “This is an example of a {label} rock.” where label is specified in the anchor images info, then the prompt text, then “This is the rock you will label.” followed by the rock image.**

images\_content = []

for img, label in base64\_anchors:

images\_content.append({"type": "text", "text": f"This is an example of a {label} rock."})

images\_content.append({"type": "image\_url", "image\_url": f"data:image/jpeg;base64,{img}"})

*# Append the main image*

images\_content.append({"type": "text", "text": f"This is the rock you will label."})

images\_content.append({"type": "image\_url", "image\_url": f"data:image/jpeg;base64,{base64\_image}"})

payload = {

"model": "gpt-4-vision-preview",

"messages": [

{

"role": "user",

"content": [

{"type": "text", "text": prompt\_text}

] + images\_content

},

],

"max\_tokens": 300,

"temperature": 0.0

}

## Lightness

anchor\_images\_info = [

("Anchors\_RGB/Low Lightness.jpg", "dark"),

("Anchors\_RGB/Medium Lightness.jpg", "medium"),

("Anchors\_RGB/High Lightness.jpg", "light")

]

'In this trial, you will rate a rock on its darkness/lightness of color. A dark rock should receive a rating of 1.00 or 2.00. A light rock should receive a rating of 8.00 or 9.00. A rock that is medium in darkness/lightness should receive a medium rating. An example of a very dark rock, a rock that is medium in darkness/lightness, and a very light rock is shown. In some cases, parts of the rock may be light and other parts may be dark. In those cases, do your best to rate the "average" lightness of the entire rock. Please try to use the full scale from 1.00 (darkest) through 9.00 (lightest) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.'**

## Grain Size

anchor\_images\_info = [

("Anchors\_RGB/Low Grain Size.jpg", "low grain size"),

("Anchors\_RGB/Medium Grain Size.jpg", "medium grain size"),

("Anchors\_RGB/High Grain Size.jpg", "high grain size")

]

'In this trial, you will rate a rock on its average grain size. Rocks with no visible grain should receive a rating of 1.00 or 2.00. Rocks with an extremely coarse and fragmented grain should receive a rating of 8.00 or 9.00. Rocks with a medium grain should receive medium ratings. An example of a rock with no visible grain, with a medium grain, and with a very coarse and fragmented grain is shown. In some cases, parts of the rock may have a fine grain and other parts may have a coarse grain. In those cases, do your best to rate the "average" grain size of the entire rock. Please try to use the full scale from 1.00 (no visible grain) through 9.00 (very coarse grain) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.**'

## Roughness

anchor\_images\_info = [

("Anchors\_RGB/Low Roughness.jpg", "low roughness"),

("Anchors\_RGB/Medium Roughness.jpg", "medium roughness"),

("Anchors\_RGB/High Roughness.jpg", "high roughness")

]

'In this trial, you will rate a rock on how smooth versus rough it appears to be. Rocks that appear to be very smooth should receive a rating of 1.00 or 2.00. Rocks that appear to be very rough should receive a rating of 8.00 or 9.00. Rocks that are medium in their smoothness/roughness should receive medium ratings. An example of a rock that is very smooth, that is medium in smoothness/roughness, and that is very rough is shown. In some cases, parts of a rock may be smooth and other parts may be rough. In those cases, do your best to rate the "average" roughness of the entire rock. Please try to use the full scale from 1.0 (smoothest) through 9.0 (roughest) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.'**

## Shininess

anchor\_images\_info = [

("Anchors\_RGB/Low Shininess.jpg", "low shininess"),

("Anchors\_RGB/Medium Shininess.jpg", "medium shininess"),

("Anchors\_RGB/High Shininess.jpg", "high shininess")

]

'An object is "shiny" if it reflects light and is glossy. Note that dark-colored objects can still be shiny. In this trial, you will rate a rock on how dull versus shiny it appears to be. Rocks that appear to be very dull should receive a rating of 1.00 or 2.00. Rocks that appear to be very shiny and glossy should receive a rating of 8.00 or 9.00. Rocks that are medium in their dullness/shininess should receive medium ratings. An example of a rock that is very dull, that is medium in dullness/shininess, and that is very shiny is shown. In some cases, parts of a rock may be dull and other parts may be shiny. In those cases, do your best to rate the "average" shininess of the entire rock. Please try to use the full scale from 1.00 (dullest) through 9.00 (shiniest) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.'**

## Organization

anchor\_images\_info = [

("Anchors\_RGB/Low Regularity.jpg", "low organization"),

("Anchors\_RGB/Medium Regularity.jpg", "medium organization"),

("Anchors\_RGB/High Regularity.jpg", "high organization")

]

'Some rocks have components that are very regular and organized, such as systematic layers, bands, or grains. Other rocks seem very disorganized, such as those with fragments that are glued together in haphazard fashion. In this trial, you will rate a rock on how disorganized versus organized it appears to be. Rocks that are very disorganized should receive a rating of 1.00 or 2.00. Rocks that are very organized should receive a rating of 8.00 or 9.00. Rocks that are medium in their organization, or that have no visible texture to rate, should receive medium ratings. An example of a rock that is very disorganized, that is medium in organization, and that is highly organized is shown. Please try to use the full scale from 1.00 (most disorganized) through 9.00 (most organized) in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.'**

## Chromaticity

anchor\_images\_info = [

("Anchors\_RGB/Warm Color.jpg", "warm color"),

("Anchors\_RGB/Cool Color2.jpg", 'cool color'),

("Anchors\_RGB/No Color2.jpg", "no color")

]

'In this trial, you will rate a rock in terms of whether it has no color, cool color, or warm color. Rocks with no color (absolute black, gray or white) should receive a rating of 1.00 or 2.00. Rocks with cool colors (blue, blue/green, and green) should receive medium ratings (4.00, 5.00, or 6.00). Rocks with very warm colors (yellow, orange, red) should receive ratings of 8.00 or 9.00. An example of a rock with no color, cool color, and warm color variation is shown. Please try to use the full scale from 1.00 (no color) through 9.00 (warmest color) in making your rating. Please try to use the full scale from 1.00 to 9.00 in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.'**

## Red/Green Hue

anchor\_images\_info = [

("Anchors\_RGB/green.jpg", "green"),

("Anchors\_RGB/neutral\_redgreen.jpg", "neutral"),

("Anchors\_RGB/red.jpg", "red")

]

'In this experiment, you will be presented with a picture of a rock. We would like you to rate the rock picture on a red-green contrast. Rocks that are most strongly red should receive ratings of 1.00 or 2.00. Rocks that are most strongly green should receive ratings of 8.00 or 9.00. Neutral rocks (black or white) that are absent of color should receive ratings of 5.00. Examples of these different cases are shown. For the remaining rocks, just decide whether the main color tends to be closer to red versus green. For example, most would agree that orange is closer to red than to green, so you might give orange rocks ratings of 2.00, 3.00, or 4.00. Likewise, most would agree that blue is closer to green than to red, so you might give blue rocks ratings of 6.00, 7.00, or 8.00. Please try to use the full scale from 1.00 to 9.00 in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.**'

## Porphyritic Texture

anchor\_images\_info = [

("Anchors\_RGB/high\_porphyritic.jpg", "high porphyritic texture"),

("Anchors\_RGB/unclear\_porphyritic.jpg", 'unclear porphyritic texture'),

("Anchors\_RGB/low\_porphyritic.jpg", "low porphyritic texture")

]

"In this experiment you will be presented with a picture of a rock. We are interested in your judgments about a very specific property of some of the rocks -- Certain kinds of rocks contain small fragments or pebbles that are glued into a separate background texture. THESE SMALL FRAGMENTS OR PEBBLES ARE SEPARATE FROM THE REST OF THE ROCK'S BACKGROUND ITSELF. We want you to rate each rock picture for this property. Rocks with no small fragments or pebbles glued into their separate background should receive a rating of 1.00 or 2.00. Rocks that definitely have small fragments or pebbles glued into their separate background should receive a rating of 8.00 or 9.00. Many rocks may be unclear cases; Some may have a coarse grain throughout, but don't really have separate small fragments glued into them. Other rocks may also be hard to judge because they have changes in shading that are not really separate glued fragments. These unclear cases should receive ratings of 4.00, 5.00 or 6.00. Examples of these different cases are shown. Please try to use the full scale from 1.00 through 9.00 in making your rating. Please try to use the full scale from 1.00 to 9.00 in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range."**

## Pegmatitic Structure

anchor\_images\_info = [

("Anchors\_RGB/strong\_pegmatite.jpg", "high pegmatitic structure"),

("Anchors\_RGB/medium\_pegmatite.jpg", 'unclear pegmatitic structure'),

("Anchors\_RGB/none\_pegmatite.jpg", "low pegmatitic structure")

]

'In this experiment you will be presented with a picture of a rock. Certain rocks have very large-sized crystals that are embedded in a SEPARATE background. The crystals will often (but not always) appear as large shiny bands. Your job in this experiment is simply to judge the extent to which the rock shown in each picture has this property. Rocks that have nothing like this property should receive ratings of 1.00 or 2.00. Rocks that have a hint of this property should receive ratings of 4.00, 5.00, or 6.00. Rocks that strongly display this property should receive ratings of 8.00 or 9.00. Examples of these different cases are shown. Please try to use the full scale from 1.00 through 9.00 in making your rating. Note: Because this property is very rare, most of the time your response will be between 1.00 and 2.00. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.'**

## Conchoidal Fracture

anchor\_images\_info = [

("Anchors\_RGB/high\_conchoidal.jpg", "high conchoidal fracture"),

("Anchors\_RGB/unclear\_conchoidal.jpg", 'unclear conchoidal fracture'),

("Anchors\_RGB/low\_conchoidal.jpg", "low conchoidal fracture")

]

'In this experiment, you will be presented with a picture of a rock. We are interested in your judgments about a very specific property of some rocks. The property is called CONCHOIDAL FRACTURES. Conchoidal fractures are formed when pieces of a brittle rock chip off and leave behind smooth, curved surfaces resembling the inside of a seashell. Conchoidal fractures are typically found in glassy or fine-grained rocks. In this trial of the experiment, you will be shown a rock picture. We want you to rate the rock picture for the extent to which it has conchoidal fractures. Rocks with flat or jagged surfaces, or rocks with no fractures should receive a rating of 1.00 or 2.00. Rocks with smooth, curved indents or fractures resembling the inside of a seashell should receive a rating of 8.00 or 9.00. Many rocks may be unclear cases: Some rocks may have fractures where pieces of the rock were chipped off, but they may not be as smooth or curved as true conchoidal fractures. Other rocks may also be hard to judge because they have changes in shading or color. These unclear cases should receive ratings of 4.00, 5.00, or 6.00. Ratings of 8.00 or 9.00 should be given only for rocks for which you are absolutely sure they have conchoidal fractures. Examples of these different cases are shown. Most rocks do not have conchoidal fractures and should receive low ratings. Please try to use the full scale from 1.00 through 9.00 in making your rating. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.**'

# CONDITION 3A: NEW PROMPTS NO ANCHOR

**Model was given the prompt text, then “This is the rock you will label.” followed by the rock image.**

images\_content = []

*# Append the main image*

images\_content.append({"type": "text", "text": f"This is the rock you will label."})

images\_content.append({"type": "image\_url", "image\_url": f"data:image/jpeg;base64,{base64\_image}"})

payload = {

"model": "gpt-4-vision-preview",

"messages": [

{

"role": "user",

"content": [

{"type": "text", "text": prompt\_text}

] + images\_content

},

],

"max\_tokens": 300,

"temperature": 0.0

}

## Organization

"Some rocks have highly organized global textures that are very regular and orderly, yielding systematic structured patterns such as stripes, bands, or physical layers. Other rocks have very disorganized global textures, such as those with fragments or crystals that seem glued together in haphazard fashion. In this trial, you will rate a rock on how disorganized or organized its global texture appears to be. A rock that has very disorganized global textures should receive ratings of 1.00 or 2.00. A rock that has very organized global textures should receive ratings of 8.00 or 9.00. A rock that is medium in its global-texture organization or that has no visible texture should receive medium ratings. Please try to use the full scale from 1.00 (most disorganized) to 9.00 (most organized) in making your ratings. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.**"

## Pegmatitic Structure

"In this experiment you will be presented with a picture of a rock. Certain rocks have long and thick shiny crystal bands that are embedded in a SEPARATE dull background. These long and thick shiny crystal bands are often (but not always) colored black or green. Your job in this experiment is to rate the extent to which the rock shown has the above-described property. Please remember that rocks with thin stripes that are not shiny do NOT have this property and that rocks that are shiny all over do NOT have this property. The rock needs to have long and thick shiny crystal bands that are embedded in a SEPARATE dull background. Rocks that have nothing like this property should receive ratings between 1.00 or 2.00. Rocks that have a hint of this property should receive ratings between 4.00, 5.00 and 6.00. Rocks that strongly display this property should receive ratings between 8.00 and 9.00. Please try to use the full scale from 1.00 through 9.00 in making your ratings. Note: Because this property is very rare, most of the time your response will be between 1.00 and 2.00. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.**"

# CONDITION 3B: NEW PROMPTS WITH ANCHOR

**Model was given “This is an example of a {label} rock.” where label is specified in the anchor images info, then the prompt text, then “This is the rock you will label.” followed by the rock image.**

images\_content = []

for img, label in base64\_anchors:

images\_content.append({"type": "text", "text": f"This is an example of a {label} rock."})

images\_content.append({"type": "image\_url", "image\_url": f"data:image/jpeg;base64,{img}"})

*# Append the main image*

images\_content.append({"type": "text", "text": f"This is the rock you will label."})

images\_content.append({"type": "image\_url", "image\_url": f"data:image/jpeg;base64,{base64\_image}"})

payload = {

"model": "gpt-4-vision-preview",

"messages": [

{

"role": "user",

"content": [

{"type": "text", "text": prompt\_text}

] + images\_content

},

],

"max\_tokens": 300,

"temperature": 0.0

}

## Organization

anchor\_images\_info = [

("Anchors\_RGB/Low Regularity.jpg", "low organization"),

("Anchors\_RGB/Medium Regularity.jpg", "medium organization"),

("Anchors\_RGB/High Regularity.jpg", "high organization")

]

"Some rocks have highly organized global textures that are very regular and orderly, yielding systematic structured patterns such as stripes, bands, or physical layers. Other rocks have very disorganized global textures, such as those with fragments or crystals that seem glued together in haphazard fashion. In this trial, you will rate a rock on how disorganized or organized its global texture appears to be. A rock that has very disorganized global textures should receive ratings of 1.00 or 2.00. A rock that has very organized global textures should receive ratings of 8.00 or 9.00. A rock that is medium in its global-texture organization or that has no visible texture should receive medium ratings. An example of a rock that has a very disorganized global texture, that is medium in its global organization, and that has a highly organized global texture is shown. Please try to use the full scale from 1.00 (most disorganized) to 9.00 (most organized) in making your ratings. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.**"

## Pegmatitic Structure

anchor\_images\_info = [

("Anchors\_RGB/strong\_pegmatite.jpg", "giant pegmatitic"),

("Anchors\_RGB/medium\_pegmatite.jpg", 'medium-coarse grained structure'),

("Anchors\_RGB/none\_pegmatite.jpg", "very fine-grained pegmatitic structure")

]

"In this experiment you will be presented with a picture of a rock. Certain rocks have long and thick shiny crystal bands that are embedded in a SEPARATE dull background. These long and thick shiny crystal bands are often (but not always) colored black or green. Your job in this experiment is to rate the extent to which the rock shown has the above-described property. Please remember that rocks with thin stripes that are not shiny do NOT have this property and that rocks that are shiny all over do NOT have this property. The rock needs to have long and thick shiny crystal bands that are embedded in a SEPARATE dull background. Rocks that have nothing like this property should receive ratings between 1.00 or 2.00. Rocks that have a hint of this property should receive ratings between 4.00, 5.00 and 6.00. Rocks that strongly display this property should receive ratings between 8.00 and 9.00. Examples of these different cases are shown. Please try to use the full scale from 1.00 through 9.00 in making your ratings. Note: Because this property is very rare, most of the time your response will be between 1.00 and 2.00. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range."**

# CONDITION 3C: NEW PROMPTS OUR RATED IMAGES

**Model was given “This is an example of a rock with a label of {label}.” where label is specified in the anchor images info, then the prompt text, then “This is the rock you will label.” followed by the rock image.**

images\_content = []

for img, label in base64\_anchors:

images\_content.append({"type": "text", "text": f"This is an example of a rock with a label of {label}."})

images\_content.append({"type": "image\_url", "image\_url": f"data:image/jpeg;base64,{img}"})

*# Append the main image*

images\_content.append({"type": "text", "text": f"This is the rock you will label."})

images\_content.append({"type": "image\_url", "image\_url": f"data:image/jpeg;base64,{base64\_image}"})

payload = {

"model": "gpt-4-vision-preview",

"messages": [

{

"role": "user",

"content": [

{"type": "text", "text": prompt\_text}

] + images\_content

},

],

"max\_tokens": 300,

"temperature": 0.0

}

## Organization

anchor\_images\_info = [

("Images\_RGB/I\_Granite\_08.jpg", "4.375"),

("Images\_RGB/M\_Gneiss\_06.jpg", "8.5"),

("Images\_RGB/M\_Schist\_04.jpg", "4.75"),

("Images\_RGB/M\_Slate\_12.jpg", "7.75"),

("Images\_RGB/S\_Breccia\_02.jpg", "1"),

("Images\_RGB/S\_Conglomerate\_08.jpg", "1.625"),

("Images\_RGB/S\_Dolomite\_12.jpg", "6"),

("Images\_RGB/S\_Rock Salt\_04.jpg", "1"),

("Images\_RGB/S\_Sandstone\_06.jpg", "8.625")

]

"Some rocks have highly organized global textures that are very regular and orderly, yielding systematic structured patterns such as stripes, bands, or physical layers. Other rocks have very disorganized global textures, such as those with fragments or crystals that seem glued together in haphazard fashion. In this trial, you will rate a rock on how disorganized or organized its global texture appears to be. A rock that has very disorganized global textures should receive ratings of 1.00 or 2.00. A rock that has very organized global textures should receive ratings of 8.00 or 9.00. A rock that is medium in its global-texture organization or that has no visible texture should receive medium ratings. Example rocks for different values on the scale 1.00-9.00 are shown. Please use these helper rocks to make your rating. Please try to use the full scale from 1.00 (most disorganized) to 9.00 (most organized) in making your ratings. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range.**"

## Pegmatitic Structure

anchor\_images\_info = [

("Images\_RGB/I\_Andesite\_10.jpg", "3.00"),

("Images\_RGB/I\_Diorite\_02.jpg", "4.00"),

("Images\_RGB/I\_Pegmatite\_07.jpg", "7.88"),

("Images\_RGB/I\_Pegmatite\_09.jpg", "8.00"),

("Images\_RGB/I\_Pegmatite\_11.jpg", "8.00"),

("Images\_RGB/M\_Schist\_11.jpg", "4.50"),

("Images\_RGB/S\_Conglomerate\_04.jpg", "4.50"),

("Images\_RGB/S\_Dolomite\_12.jpg", "1.00"),

("Images\_RGB/S\_Sandstone\_07.jpg", "1.00"),

]

"In this experiment you will be presented with a picture of a rock. Certain rocks have long and thick shiny crystal bands that are embedded in a SEPARATE dull background. These long and thick shiny crystal bands are often (but not always) colored black or green. Your job in this experiment is to rate the extent to which the rock shown has the above-described property. Please remember that rocks with thin stripes that are not shiny do NOT have this property and that rocks that are shiny all over do NOT have this property. The rock needs to have long and thick shiny crystal bands that are embedded in a SEPARATE dull background. Rocks that have nothing like this property should receive ratings between 1.00 or 2.00. Rocks that have a hint of this property should receive ratings between 4.00, 5.00 and 6.00. Rocks that strongly display this property should receive ratings between 8.00 and 9.00. Example rocks for different values on the scale 1.00-9.00 are shown. Please use these helper rocks to make your rating. Please try to use the full scale from 1.00 through 9.00 in making your ratings. Note: Because this property is very rare, most of the time your response will be between 1.00 and 2.00. **Please respond ONLY with a continuous numeric decimal value, allowing for any decimal places within the range of 1.00 to 9.00. Precision is key, and values should NOT be constrained to 0.05 increments. Your response can include any decimal point to the hundredths place within the specified range."**